

Supplementary Information: Rank fidelity determines when certified optimisation is safe

Sehaj Singh

August 2026

Full Proof of Theorem 1 (Rank-Fidelity Phase Transition)

S.1 Setup

Let F_θ be a smooth parametric family with $\theta \in \mathbb{R}^d$. Let $T : \mathbb{R}^d \rightarrow \mathbb{R}$ be the truth functional and $C : \mathbb{R}^d \rightarrow \mathbb{R}$ be the certificate with $C(\theta) \geq T(\theta)$ for all θ .

Define the slack $S(\theta) = C(\theta) - T(\theta) \geq 0$.

We work under three assumptions:

1. **(Gaussian truth)** $T(\theta)$ is approximately Gaussian over F_θ under a reference distribution $p(\theta)$.
2. **(Approximate independence)** S and T are approximately uncorrelated: $\text{Cov}(S, T) \approx 0$.
3. **(Positive slack)** $S(\theta) \geq 0$ for all θ .

S.2 Proof

Proof. Write $C = T + S$ where $S \geq 0$. We wish to compute $\rho_S(C, T) = \rho_S(T + S, T)$.

Step 1: Exact Pearson transition (distribution-free). Since $C = T + S$ with $S \perp T$ and both of finite variance, and requiring *only* independence (no Gaussianity):

$$\begin{aligned} \text{Cov}(C, T) &= \text{Cov}(T + S, T) = \text{Var}(T) = \sigma_T^2 \\ \text{Var}(C) &= \sigma_T^2 + \sigma_S^2 = \sigma_T^2 (1 + r), \quad r := \sigma_S^2 / \sigma_T^2 \\ \rho_P(C, T) &= \frac{\sigma_T^2}{\sqrt{\sigma_T^2 (1 + r)} \sigma_T} = \frac{1}{\sqrt{1 + r}}, \quad \rho_P(1) = \frac{1}{\sqrt{2}} \approx 0.707. \end{aligned}$$

This is exact and makes no distributional assumption on S or T beyond finite second moments; it is verified against Monte Carlo to 3.6×10^{-4} .

Step 2: Exact Spearman for the canonical non-negative-slack model. The physically relevant slack is non-negative ($S \geq 0$, a certificate upper-bounds the truth). For the reference model $T \sim \mathcal{N}(0, 1)$, $S \sim \text{HalfNormal}(\sigma)$ with $\sigma = \sqrt{r/(1 - 2/\pi)}$, the Spearman coefficient is

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

where $F_S(x) = \text{erf}(x/(\sigma\sqrt{2}))$ for $x \geq 0$ (and 0 otherwise), F_T is the standard-normal CDF, and $C = T + S$. Both expectations are evaluated to machine precision (10^{-10}) by a 64×64 Gauss-Hermite product rule. This yields $\rho_S(0.1) = 0.957$, $\rho_S(1) = 0.716$, $\rho_S(10) = 0.334$. As $r \rightarrow \infty$, an order-statistic expansion gives the universal asymptote

$$\rho_S(r) \sim \kappa r^{-1/2}, \quad \kappa \approx 2.14.$$

Step 3: Analytical design rule $g(r)$. Projecting the exact curve onto the compact family $g(r) = 1 - \alpha \text{ReLU}(\tanh(\beta \ln r))$ gives, analytically (zero sampling noise):

$$\alpha = 0.812, \quad \beta = 0.609,$$

with mean absolute error 0.067 over $r \in [10^{-2}, 10^2]$; the largest error (0.28) sits at the transition $r \approx 1$, where the true curve is smooth but $g(r)$ is piecewise linear in $\ln r$. Practitioners should consult the exact curve; $g(r)$ is a compact design-rule summary. The ReLU enforces $g(r) = 1$ for $r \leq 1$ and preserves the empirical asymmetry that rank fidelity never exceeds the no-slack value. $\square$

Full Proof of Theorem 2 (Dependent Slack)

When $\text{Cov}(S, T) = \rho_{ST}\sigma_S\sigma_T \neq 0$:

$$\text{Cov}(C, T) = \text{Var}(T) + \text{Cov}(S, T) = \sigma_T^2 + \rho_{ST}\sigma_S\sigma_T$$

$$\text{Var}(C) = \sigma_T^2 + \sigma_S^2 + 2\rho_{ST}\sigma_S\sigma_T$$

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

This is the exact Pearson identity for correlated slack, verified against correlated-Gaussian Monte Carlo (300k samples) to 1.3×10^{-3} . Since $\text{Var}(C) = \sigma_T^2(1 + r_\rho)$, the independent-slack formula applied to the signed effective ratio $r_\rho = r + 2\rho_{ST}\sqrt{r}$ returns the identical ρ_P ; the ReLU-tanh generalisation gives $\rho_S \approx g(r_\rho)$.

The critical threshold $r_c(\rho_{ST})$, defined by $\rho_S(r_\rho(r_c)) = 0.5$, shifts continuously with ρ_{ST} : for $\rho_{ST} < 0$, $r_\rho(r) < r$, expanding the safe regime; for $\rho_{ST} > 0$, it contracts it. This precisely explains the tokamak equilibrium result (below), where the truncation slack is strongly anti-correlated with truth ($\rho_{ST} = -0.81$) yet $r = 0.017 \ll 1$ leaves fidelity at $\rho_S = 0.993$.

Additional experimental detail

Free-boundary Grad-Shafranov equilibrium (real solver)

Tokamak equilibrium is computed with the open-source FreeGS solver (v0.8.2), a free-boundary Grad-Shafranov equilibrium solver of the same class used in EFIT/CHEASE-style reconstruction pipelines. The fixed-boundary domain is the TestTokamak machine ($R \in [0.1, 2.0]$ m, $Z \in [-1, 1]$ m); free-boundary shape control constrains two X-points at $R = 1.1$ m, $Z = \pm z_x$. Twelve configurations draw plasma current $I_p \sim 10^{U(4.90, 5.48)} = [0.8, 300]$ kA, pressure on axis $p_0 \sim 10^{U(2.78, 3.34)} = [600, 2200]$ Pa, and shaping $z_x \sim U[0.40, 0.80]$ m. For each, the certificate C is the stored thermal energy w_{th} of a truncated 17×17 Picard solve (4 iterations); the truth T is w_{th} of the converged 65×65 solve (up to 40 Picard iterations). Coarse solves are 15–25 $\times$ cheaper. Spearman $\rho_S = 0.993$ ($p < 10^{-10}$), slack ratio $r = 0.017$, slack–truth correlation $\rho_{ST} = -0.81$; all 12 solves converge on both grids. Reproduced by `tokamak_gs_real.py`; results in `results/tokamak_results.json`.

HPC-scale Oseen (linearised Navier–Stokes) AMG

The Oseen system $[-\nu\Delta + \mathbf{c} \cdot \nabla, 0; 0, -\Delta][u; p]^\top = [f; g]^\top$ is discretised by cell-centred finite volumes with two unknowns per cell on 128^3 (4.19M unknowns, 12 configurations) and 256^3 (33.5M unknowns, 3 configurations) grids, assembled in sparse CSR. Viscosity ν spans $[10^{-3}, 10]$; convection $\mathbf{c}$ spans $[10^{-2}, 10^2]$. The certificate is a spectral (Gershgorin upper) bound from 25 power iterations; the truth

is the measured Jacobi smoothing convergence factor over 20 sweeps. The slack ratio $r \approx 10^7 \gg 1$ places solver selection in the overflow (untrustworthy) regime: Spearman $\rho_S = 0.38$ ($p = 0.23$), indistinguishable from zero and reproduced at 256³. Reproduced by `navier_stokes_scale.py`; results in `results/navier_stokes_results.json`.

CIFAR-10 certified robustness

The CIFAR-10 rank-fidelity experiment (52 MLPs) and the Goodhart sweep (8 models) were executed on Kaggle kernel `sehajrsingh/rank-fidelity-cpu-v10` (CPU, 12-hour limit). Each model trains on a 6,000-image CIFAR-10 subset for 10 epochs with the IBP weight and training perturbation ramped linearly over the first half of training. Per-model certified accuracy (exact interval bound propagation through the sequential linear-ReLU stack) and PGD robust accuracy (10 steps, $\alpha = 0.008$, $\epsilon = 8/255$) are recorded in `results/cifar10_verified_results.json`; the summary statistic is Spearman $\rho_S = 0.368$ ($p = 0.007$), $n = 52$.

References