THE SIGMOID STRIKES BACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Physics-informed objectives for second-order equations read a network's second input derivatives. For the model reaction–diffusion equation, we prove that the risk over piecewise-linear networks has infimum zero, yet every minimizing sequence converges in L^2 to the source rather than the true solution. Its residual vanishes while its solution error approaches the norm of the solution's Laplacian; for unit-norm high-frequency sources, that error approaches one as the solution norm vanishes. The same order reduction holds at every depth and for general elliptic and fully nonlinear operators, with wrong-limit conclusions under the stated solvability conditions. The regression identity is exact at collocation points, and output-weight gradient descent with fixed piecewise-linear features reaches the spurious limit as their spans become dense. An almost-everywhere Hessian-only penalty is constant on this class. Piecewise-linear energy models also collapse under score matching to a zero score. In contrast, smooth saturating networks approximate sufficiently regular solutions together with their second derivatives. When the residual and boundary misfit control solution error, every minimizing sequence is consistent; logistic sigmoid and hyperbolic tangent networks achieve a quantitative rate.

1 INTRODUCTION

The default activation in contemporary deep learning is piecewise linear, as in the rectified linear unit and its variants, or piecewise polynomial, as in the rectified power units. The logistic sigmoid and the hyperbolic tangent are treated as relics, blamed for vanishing gradients and slow optimization. This judgment reverses for objectives that depend on derivatives of the network with respect to its *input*, rather than on its output alone. The leading example is the physics-informed neural network (Lagaris et al., 1998; Raissi et al., 2019; Karniadakis et al., 2021), which fits the residual of a partial differential equation by differentiating the network through automatic differentiation. Score-based generative models, Hamiltonian and Lagrangian networks, and the Deep Ritz method (E & Yu, 2018) share this dependence on input derivatives. The rectified linear unit has zero classical curvature wherever it is differentiable; the hyperbolic tangent does not.

The decisive quantity is the order of differentiation required by the objective, measured against the activation's smoothness. This paper is about the objectives that read a *second* input-derivative: the physics-informed residual of a second-order equation, and the score-matching objective of generative modeling. We show that on these objectives the failure is categorical rather than gradual.

1.1 OUR RESULTS

We work on a bounded domain $\Omega \subset \mathbb{R}^d$ with a uniformly elliptic second-order operator $\mathcal{L}$ and a source f. A network u is scored by the physics-informed risk

$$J(u) = \int_\Omega (\mathcal{L}u - f)^2 dx + \mu \, \|u - g\|^2_{L^2(\partial\Omega)} \, .$$

Its first term is the squared residual of the equation over Ω. Its second term is a weight $\mu \geq 0$ times the squared L^2 misfit between u and the Dirichlet data g on the boundary $\partial\Omega$. Both terms are made precise in Theorem 2.4. The *interior risk* is the case $\mu = 0$, and $\|\cdot\|$ without a subscript is the norm of $L^2(\Omega)$. Our central finding is a categorical separation on this risk between piecewise-linear and smooth saturating activations, stated first for the model problem $-\Delta u^\star + u^\star = f$, in which Δ is

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.

the Laplacian and $u^\star$ the true solution. Each main result is previewed here, with the body statement that restates it named beside it, and every derivative of a network in these previews is read at the almost-everywhere values automatic differentiation returns (Theorem 2.2).

The Collapse. We first establish this collapse for the model problem.

Theorem 1.1 (Inconsistency of piecewise-linear solvers; restated as Theorem 3.2)**.** *Consider the model problem with a non-harmonic solution $u^\star$, and let $\mathcal{N}_{\mathsf{PL}}$ be the class of piecewise-linear networks. Over this class the physics-informed risk J has infimum zero. Yet every minimizing sequence (u_n) converges in $L^2(\Omega)$ to the source f rather than to $u^\star$. Consequently $\|u_n - u^\star\| \to \|\Delta u^\star\| =: \delta(f) > 0$ while the training residual vanishes. Moreover, over unit-norm sources $\delta(f)$ can approach $\|f\| = 1$ while $\|u^\star\| \to 0$.*

The reason is one line: a piecewise-linear network is affine on each cell of its partition, so its Hessian vanishes off the cell boundaries and $-\Delta u + u$ is read as u. The risk is then the least-squares misfit $\int_\Omega (u - f)^2 dx$, whose minimizers fit the source rather than the solution. Nothing about the model operator is responsible.

Theorem 1.2 (General elliptic inconsistency; restated as Theorem 4.1)**.** *Let $\mathcal{L}u = -\operatorname{div}(A\nabla u) + b \cdot \nabla u + cu$ be uniformly elliptic. Here ∇ is the gradient and* div *the divergence, A is the coefficient matrix, b the first-order coefficient and c the zeroth-order coefficient. The reduced first-order operator of $\mathcal{L}$ is $\tilde{\mathcal{L}}u = \tilde{b} \cdot \nabla u + cu$, with $\tilde{b} = b - \nabla \cdot A$: the operator left once the Hessian is annihilated. Without drift, that is when $\tilde{b} \equiv 0$, the collapse holds verbatim with limit f/c. In that case the infimum of J over $\mathcal{N}_{\mathsf{PL}}$ is zero for every $\mu \geq 0$. Every minimizing sequence converges in $L^2(\Omega)$ to f/c, which differs from $u^\star$ unless the principal part annihilates $u^\star$. With drift, suppose $\tilde{b}$ is Lipschitz and $\tilde{\mathcal{L}}w = f$ is solvable in $H^1(\Omega)$. Then the interior risk still has infimum zero on $\mathcal{N}_{\mathsf{PL}}$, and every L^2-limit point $\tilde{u}$ of a minimizing sequence solves $\tilde{\mathcal{L}}\tilde{u} = f$ in the sense of distributions. Such a limit point misses $u^\star$ unless the principal part annihilates $u^\star$.*

The same reduction reaches fully nonlinear operators, where the Hessian may enter nonlinearly.

Theorem 1.3 (Nonlinear collapse; restated as Theorem 4.2)**.** *Let $\mathcal{F}[u] = F\left(\nabla^2 u, \nabla u, u, x\right)$ be an operator given by a function F of the Hessian $\nabla^2 u$, the gradient, the value and the point, with F continuous in its Hessian argument at the origin and the residual square integrable on the networks considered. Its reduced first-order part is $\mathcal{F}_0[u] = F(0, \nabla u, u, x)$. Write $J_{\mathcal{F}}(u) = \|\mathcal{F}[u] - f\|^2 + \mu B(u)$ for its risk. On $\mathcal{N}_{\mathsf{PL}}$ the interior residual is $\|\mathcal{F}_0[u] - f\|^2$, and two regimes occur. Where $\mathcal{F}_0[u]$ reads neither ∇u nor u, as for Monge–Ampère and the p-Laplacian with $p \geq 2$, that residual is a constant and the equation gives no interior training signal. In the other regime, take the interior risk, let $F(0, p, z, x)$ be Lipschitz in (p, z) uniformly in x, and let $\mathcal{F}_0[w] = f$ be solvable in $H^1(\Omega)$. Then the infimum of $J_{\mathcal{F}}$ is zero, and every H^1-limit point $\tilde{u}$ of a minimizing sequence for $J_{\mathcal{F}}$ solves $\mathcal{F}_0[\tilde{u}] = f$. For a twice-differentiable true solution, such a limit point misses $u^\star$ unless $\mathcal{F}[u^\star] = \mathcal{F}_0[u^\star]$.*

Repairs. A network's wall is the lowest order at which every derivative of it vanishes. Raising the activation's degree moves the wall; at degree one, the piecewise-linear case, no depth moves it, and penalizing the objective does not reach the curvature at all.

Proposition 1.4 (Depth, degree, and the wall location; restated as Theorem 4.3)**.** *A depth-L network u with activations piecewise polynomial of degree at most p is cellwise of degree at most p^L. So every derivative $D^\alpha u$ of order $|\alpha| > p^L$ vanishes almost everywhere. For $p = 1$ the wall stands at order at most two at every depth; for $p \geq 2$, a nonzero derivative of order r requires depth $L \geq \log_p r$.*

Penalizing the curvature is the other repair one might try.

Theorem 1.5 (Autodiff regularization cannot read piecewise-linear curvature; restated as Theorem 5.1)**.** *Let $R[u] = \int_\Omega r\left(\nabla^2 u, \nabla u, u, x\right) dx$ be read at the almost-everywhere derivatives and finite on $\mathcal{N}_{\mathsf{PL}}$, and let $J_\lambda = J + \lambda R$. Then $R[u] = \int_\Omega r(0, \nabla u, u, x)\, dx$ for every $u \in \mathcal{N}_{\mathsf{PL}}$, so no regularizer reads the curvature of the class. In particular every regularizer $r\left(\nabla^2 u, x\right)$ is a constant R_0 on $\mathcal{N}_{\mathsf{PL}}$. Adding it to either J or $J_{\mathcal{F}}$ preserves the corresponding minimizing sequences and limits in Theorems 1.1 to 1.3, and shifts each infimum by λR_0.*

Score Matching. Score matching fits a model density $p_\theta \propto e^{\varphi_\theta}$, given by an energy φ_θ, to a data density q. It does so by matching the model's score $\nabla\varphi_\theta$ to the score of q. Hyvärinen's identity writes the objective through the divergence of the score, a second derivative of the energy, and the same annihilation removes that term.

Theorem 1.6 (Score-matching collapse; restated as Theorem 4.5)**.** *Let the energy be $\varphi_\theta \in \mathcal{N}_{\mathsf{PL}}$ and the objective be read, as in training, at the almost-everywhere derivatives. Then $\Delta\varphi_\theta = 0$ almost everywhere, the objective reduces to $\mathbb{E}_{x\sim q}\left[\frac{1}{2}\left|\nabla\varphi_\theta\right|^2\right]$, and its infimum over $\mathcal{N}_{\mathsf{PL}}$ is zero, attained by the constant energy, whose score $s_\theta = \nabla\varphi_\theta$ is identically zero. For non-uniform q the Fisher divergence there is positive, so the procedure is inconsistent.*

What Training Reaches. The statements so far concern exact integrals and infima; the sample objective and the trajectory reach the same object.

Proposition 1.7 (Exactness under collocation; restated as Theorem 5.2)**.** *At any collocation points, evaluate the model operator as training does. Then a piecewise-linear network's empirical physics-informed risk equals its empirical regression misfit to f. The identity holds exactly at the sample points, not up to quadrature error, and interpolating f drives the empirical risk to zero.*

Output-weight training reaches the same spurious limit.

Theorem 1.8 (Gradient descent reaches the spurious limit; restated as Theorem 5.3)**.** *Fix piecewise-linear features and train the output weights alone. For the model operator the risk is a convex quadratic in the weights, and gradient flow, like gradient descent with a small enough step, converges to a global minimizer. For the interior risk the trained network is the $L^2(\Omega)$-orthogonal projection of f onto the feature span, and otherwise it minimizes the risk over the span. It tends to f as the spans become dense in $L^2(\Omega)$, or in $H^1(\Omega)$ when $\mu > 0$.*

The Smooth Side. Smooth saturating activations instead give consistent solvers, with a quantitative rate.

Theorem 1.9 (Consistency of smooth solvers, with rate; restated as Theorem 6.1)**.** *Let $\sigma \in \mathsf{S}$ be a smooth saturating activation, let $\mathcal{N}_{\mathsf{S}}$ be the class of networks with such activations, and let the solution $u^\star$ extend to a function on $\mathbb{R}^d$ with $s \geq 2$ continuous derivatives. Then $\inf_{u\in\mathcal{N}_{\mathsf{S}}} J(u) = 0$, and for the model problem with a boundary term every minimizing sequence converges to $u^\star$ in L^2. Take the logistic sigmoid or the hyperbolic tangent, $s > 2$ and $\varepsilon > 0$. Then there are networks with N nonzero weights and $J = O\big(N^{-2(s-2-\varepsilon)/d}\big)$. For the model problem with a boundary term, its L^2 error is $O\big(N^{-(s-2-\varepsilon)/d}\big)$.*

The inconsistency theorem, its extensions and its score-matching analogue are to our knowledge new; the curvature obstruction underlying Theorem 1.4 is known (Leng & Thiyagalingam, 2022) and the rate in Theorem 1.9 is imported (Gühring & Raslan, 2021), and we mark the provenance of every statement at its point of use.

1.2 Our Techniques

The proofs combine three devices. Because the reduced functional is a least-squares misfit to f, what was known as an approximation floor becomes a wrong limit that training reaches at zero loss. Reading the training convention literally makes this reduction exact at every collocation point, and fixing the features makes the objective a convex quadratic with a closed-form limit.

1.3 Other Related Work

The score-matching objective and the identity that removes the normalizing constant are due to Hyvärinen (2005). Shen et al. (2023) use differentiable rectified-power networks for directly parametrized score estimation and argue that rectified-linear networks are unsuitable when derivatives enter the objective. Under our scalar-energy parametrization the objective reads one further derivative, and its second-order term is annihilated outright (Theorem 4.5).

The Curvature Obstruction. Leng & Thiyagalingam (2022) prove that a network with rectified-linear-like activations has a vanishing Hessian and so cannot form a permissible function space for

second-order equations. Theorem 3.2 sharpens their obstruction into a statement about minimizing sequences and limits: an approximation floor warns that the target is unreachable, whereas an inconsistency theorem shows the procedure reporting success at the wrong object.

That minimizing a physics-informed loss need not control the solution has been observed in three other settings. Wang et al. (2022) construct, for a Hamilton–Jacobi–Bellman operator, functions whose L^p residual is arbitrarily small while the solution error is arbitrarily large, an instability of the operator in the chosen norm rather than a property of any network class. Berrone et al. (2022) exhibit a rectified-linear network driving a *variational* residual to zero at a spurious solution when the inf-sup condition fails. Closest of all, Doumèche et al. (2025) give minimizing sequences whose physics-informed risk tends to zero and which converge to an explicit wrong limit; their mechanism is a loss of coercivity, with the H^1 norms diverging, and their repair is a Sobolev penalty. Ours is the complementary case: the operator rather than the norm degenerates, the wrong limit is the source f, and the second-order part of their Sobolev penalty is, by Theorem 5.1, constant on the class.

1.4 Discussion

Automatic differentiation agrees with classical derivatives almost everywhere for piecewise-analytic programs (Lee et al., 2020), and under Theorem 2.2 it returns a zero Hessian even at the interfaces of a piecewise-linear network. The distributional Hessian can carry mass on those interfaces (Theorem B.4), which the pointwise residual does not evaluate. For smooth-network PDE solvers, automatic differentiation can improve training relative to finite differences (Chen et al., 2025); the failure we isolate occurs at the piecewise-linear kinks. The separation is therefore between the piecewise-linear activations and the smooth ones; we claim no separation among the smooth ones. The separation turns on the vanishing almost-everywhere second derivative, not on saturation, and consistency holds for the sigmoid linear unit and the Gaussian error linear unit as well (Theorem D.2).

The open problems of all-weight descent and trained weak forms are stated in Section E.1.

1.5 Paper Organization

Section 2 fixes notation and collects the facts on approximation, piecewise-linear derivatives, and elliptic regularity. Section 3 proves the inconsistency theorem. Section 4 extends the result to general and nonlinear operators and to score matching, and identifies the barrier for each activation degree. Section 5 examines the optimization objective, while Section 6 proves consistency with a rate.

2 Preliminaries

2.1 Domains and Networks

We first fix the domain and the function spaces in which the statements are measured.

Definition 2.1 (Domain and function spaces)**.** Throughout, $\Omega \subset \mathbb{R}^d$ is a bounded domain whose boundary is of class C^2, except in Theorem A.1, where $\Omega = \mathbb{T}^d = (\mathbb{R}/2\pi\mathbb{Z})^d$ is the flat torus. We write $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ for the norm and inner product of $L^2(\Omega)$, and $\|\cdot\|_{L^2(\partial\Omega)}$ for the norm of the trace on $\partial\Omega$. The Sobolev spaces are $H^k(\Omega) = W^{k,2}(\Omega)$ and $W^{k,\infty}(\Omega)$, the latter normed by $\|u\|_{W^{k,\infty}(\Omega)} = \max_{|\alpha|\le k} \|D^\alpha u\|_{L^\infty(\Omega)}$, where α is a multi-index and D^α the weak derivative. The regularity of $\partial\Omega$ enters only through the trace theorem, the density of smooth functions and H^2 regularity, each from Evans (2010).

A network with a piecewise-polynomial activation is smooth only cell by cell, so its derivatives need a convention; we fix the one training uses.

Definition 2.2 (Almost-everywhere derivative of a cellwise-smooth map)**.** Let u be continuous on $\overline{\Omega}$ and smooth on the interior of each cell of a finite partition of Ω whose interfaces form a Lebesgue-null set. The cells are cut out of Ω by polytopes for a piecewise-linear network and by semialgebraic sets for a piecewise-polynomial one (Theorem 4.3). For a multi-index α the *almost-everywhere derivative* $D^\alpha u$ is the classical derivative on the union of the cell interiors, and $D^\alpha u = 0$ *almost everywhere* means that it vanishes there. This is the value automatic differentiation and collocation

return off the interfaces: the second derivative of a rectified linear unit is zero, and this is the convention under which physics-informed and score-matching objectives are evaluated in training. On an interface, automatic differentiation implements the derivative of each piecewise-linear activation as a piecewise-constant function whose own derivative it takes to be zero, so it returns a zero Hessian for a piecewise-linear network there as well. Every derivative of such a network below is read this way. For $|\alpha| \leq 1$ it is the weak derivative of Theorem 2.1, but the almost-everywhere Hessian of a piecewise-linear network with a kink is not its distributional Hessian, which has in addition a part on the interfaces (Theorem B.4).

The networks, and the activation classes the results compare, are the following.

Definition 2.3 (Feedforward network and activation classes)**.** A feedforward network with activation $\sigma\colon \mathbb{R} \to \mathbb{R}$ applied entrywise, depth L and widths at most W computes $u(x) = A_L\sigma\left(A_{L-1}\sigma\left(\cdots\sigma\left(A_0x + b_0\right)\cdots\right) + b_{L-1}\right) + b_L$. We distinguish three activation classes. The class PL is *piecewise linear*: σ is continuous and affine on finitely many intervals. Examples are the rectified linear unit $\rho(t) = \max\{t, 0\}$ and its leaky and clipped variants; a PL network is a continuous piecewise-linear map on a finite polytopal partition. Next, P_p is *piecewise polynomial* of degree at most p, so $\mathsf{PL} = \mathsf{P}_1$, and the degree-$k$ rectified power unit lies in P_k. Finally, S is *smooth saturating*: $\sigma \in C^\infty(\mathbb{R})$ is non-polynomial and bounded with every derivative, as the logistic sigmoid and $\tanh$ are. We write $\mathcal{N}_{\mathsf{X}}$ for the maps $\Omega \to \mathbb{R}$ realizable with activation class X at any finite depth and width.

2.2 The Physics-Informed Risk and Imported Facts

The objective of the physics-informed results is the following risk.

Definition 2.4 (Operator and risk)**.** Let $\mathcal{L}u = -\operatorname{div}(A\nabla u) + b\cdot\nabla u + cu$ be a second-order linear operator. Its matrix $A = A(x)$ is uniformly elliptic, meaning $A \succeq \lambda I$ for some $\lambda > 0$. It is also Lipschitz, so that its row divergence $\nabla\cdot A$ is bounded. The coefficients b and c are bounded, and $c \geq c_0$ for a constant $c_0 > 0$. Given $f \in L^2(\Omega)$ and Dirichlet data g that is the trace on $\partial\Omega$ of some $w_g \in H^1(\Omega)$, the *physics-informed risk* is

$$J(u) = \int_\Omega \left(\mathcal{L}u(x) - f(x)\right)^2 dx + \mu B(u), \qquad B(u) = \|u - g\|^2_{L^2(\partial\Omega)}, \qquad \mu \geq 0, \tag{1}$$

the boundary term reading the trace of u, and the *model problem* is

$$\mathcal{L} = -\Delta + 1, \qquad -\Delta u^\star + u^\star = f \text{ in } \Omega, \qquad u^\star = g \text{ on } \partial\Omega, \tag{2}$$

whose weak solution $u^\star \in H^1(\Omega)$ exists and is unique by Theorem 2.7.

The following facts are drawn from the literature; we invoke them and never reprove them.

Fact 2.5 (Derivative density and rate; Hornik, 1991; Gühring & Raslan, 2021)**.** *Let $\sigma \in \mathsf{S}$.*

(i) *For every $k \geq 0$, every $G \in C^k(\mathbb{R}^d)$ and every $\eta > 0$ there is a network $u(x) = \sum_{l=1}^m a_l\sigma\left(w_l\cdot x + \beta_l\right)$ in $\mathcal{N}_{\mathsf{S}}$, with one hidden layer, activation σ and no output bias, such that $\|u - G\|_{W^{k,\infty}(\Omega)} \leq \eta$ (Hornik, 1991, Theorem 3).*

(ii) *Let σ be the logistic sigmoid or the hyperbolic tangent, let $Q \supset \overline{\Omega}$ be a cube, and let $s > k \geq 1$ be integers. Fix $\varepsilon > 0$. There are constants C and N_0, depending only on σ, the dimension, the two orders, ε and the cube, with the following property. Let G lie in the unit ball of $W^{s,\infty}(Q)$, and let $N \geq N_0$. Then some network u with at most N nonzero weights has $\|u - G\|_{W^{k,\infty}(\Omega)} \leq CN^{-(s-k-\varepsilon)/d}$ (Gühring & Raslan, 2021, Theorem 4.9 and Table 1, on $(0,1)^d$, carried to Q by an affine change of variables).*

The next fact is the observation the collapse starts from.

Fact 2.6 (Vanishing Hessian of a piecewise-linear network; Leng & Thiyagalingam, 2022)**.** *Every $u \in \mathcal{N}_{\mathsf{PL}}$ is continuous piecewise linear, so its input-Hessian vanishes almost everywhere in the sense of Theorem 2.2; this is immediate from the cellwise-affine structure and is the obstruction proved by Leng & Thiyagalingam (2022).*

For the model problem we also need that the solution exists and, for zero boundary data, has two derivatives in L^2.

Fact 2.7 (Well-posedness and H^2 regularity of the model problem; Evans, 2010)**.** *For $f \in L^2(\Omega)$ and g the trace of some $w_g \in H^1(\Omega)$, the model problem of Eq.* (2) *has a unique weak solution $u^\star \in H^1(\Omega)$, by the Lax–Milgram theorem applied to $u^\star - w_g \in H^1_0(\Omega)$ (Evans, 2010, Section 6.2, Theorem 1). If $g = 0$, then $u^\star$ lies in $H^2(\Omega)$. Its norm there obeys $\|u^\star\|_{H^2(\Omega)} \le C\,(\|f\| + \|u^\star\|)$, with C depending only on Ω (Evans, 2010, Section 6.3, Theorem 4, which uses that $\partial\Omega$ is C^2).*

Finally, the collapse proofs reach their infimum by approximating from the piecewise-linear class in $L^2(\Omega)$ and in $H^1(\Omega)$.

Lemma 2.8 (Density of the piecewise-linear class; proved in Section A)**.** *The class $\mathcal{N}_{\mathsf{PL}}$ is dense in $C(\overline{\Omega})$, hence in $L^2(\Omega)$, and dense in $H^1(\Omega)$.*

3 The Inconsistency Theorem

Everything below rests on one computation, stated first.

Lemma 3.1 (Order reduction; proved in Section A)**.** *For every $u \in \mathcal{N}_{\mathsf{PL}}$, almost everywhere on Ω,*

$$\mathcal{L}u = \tilde{b} \cdot \nabla u + cu, \qquad \tilde{b} = b - (\nabla \cdot A)\,,$$

so the entire principal part is annihilated. For the model operator this reads $\mathcal{L}u = u$ almost everywhere.

For the model operator the reduction turns the interior residual into a regression onto the source.

Theorem 3.2 (Loss collapse and inconsistency; restatement of Theorem 1.1; proved in Section A)**.** *Consider the risk of Eq.* (1) *for the model problem of Eq.* (2)*, with any $\mu \ge 0$.*

(a) Collapse. *For every $u \in \mathcal{N}_{\mathsf{PL}}$ the interior residual equals $\int_\Omega (u-f)^2\,dx$, hence $J(u) \ge \int_\Omega (u-f)^2\,dx$.*

(b) Zero infimum. $\inf_{u \in \mathcal{N}_{\mathsf{PL}}} J(u) = 0$.

(c) Wrong limit. *Every minimizing sequence (u_n) converges to f in $L^2(\Omega)$, and consequently*

$$\lim_{n\to\infty} \|u_n - u^\star\| = \|f - u^\star\| = \|\Delta u^\star\| =: \delta(f) > 0$$

whenever $u^\star$ is non-harmonic, while the training residual tends to zero.

(d) Error as large as the signal. *For every $\varepsilon > 0$ there is a source with $\|f\| = 1$, $\delta(f) \ge 1 - \varepsilon$, and $\|u^\star\| \le \varepsilon$; in particular the error relative to the true solution, $\delta(f)/\|u^\star\|$, is unbounded.*

Parts (a) and (c) bound the interior term alone, so they hold for every nonnegative boundary functional in place of B; part (b) is where the boundary data enters.

On the flat torus a family of sources like that of part (d) is explicit, and Theorem A.1 in Section A computes $\delta(f)$ frequency by frequency.

4 Generality and Transfer of the Collapse

The inconsistency of Section 3 is not an artifact of the model equation: the order reduction applies to linear elliptic operators with drift, to fully nonlinear operators and to score matching. Theorem 4.3 then locates the wall of depth-L piecewise-polynomial networks.

4.1 General and Nonlinear Operators

Without drift the spurious limit is explicit. With drift, a first-order equation carries boundary data only where the drift enters the domain, so the Dirichlet term and the reduced residual can conflict (Theorem B.1). Part (ii) of the theorem below is therefore stated for the interior risk.

Theorem 4.1 (General elliptic inconsistency; restatement of Theorem 1.2; proved in Section B)**.** *Let $\mathcal{L}$ be as in Theorem 2.4, with reduced operator $\tilde{\mathcal{L}}u = \tilde{b} \cdot \nabla u + cu$, where $\tilde{b} = b - (\nabla \cdot A)$. Let $u^\star \in H^1(\Omega)$ solve $\mathcal{L}u^\star = f$ with the boundary data g.*

(i) No drift. *If $\tilde{b} \equiv 0$, that is if $\mathcal{L} = -\sum_{i,j} A_{ij}\partial_i\partial_j + c$, then for every $\mu \geq 0$ one has $\inf_{u \in \mathcal{N}_{\mathsf{PL}}} J(u) = 0$, and every minimizing sequence converges in $L^2(\Omega)$ to $\tilde{u} = f/c$, whose distance $\|f/c - u^\star\|$ from $u^\star$ is positive unless $f = cu^\star$, that is unless the second-order part $-\sum_{i,j} A_{ij}\partial_i\partial_j u^\star$ vanishes.*

(ii) Drift. *If $\tilde{b}$ is Lipschitz, $\mu = 0$, and $\tilde{\mathcal{L}}w = f$ has a solution in $H^1(\Omega)$, then $\inf_{u \in \mathcal{N}_{\mathsf{PL}}} J(u) = 0$, and every L^2-limit point $\tilde{u}$ of a minimizing sequence solves the reduced equation $\tilde{\mathcal{L}}\tilde{u} = f$ in the sense of distributions; such a limit point equals $u^\star$ only if $\tilde{\mathcal{L}}u^\star = \mathcal{L}u^\star$, that is only if the second-order part of $\mathcal{L}$ annihilates $u^\star$.*

The collapse also reaches fully nonlinear problems, where it takes one of two forms.

Theorem 4.2 (Nonlinear collapse; restatement of Theorem 1.3; proved in Section B)**.** *Let $\mathcal{F}[u](x) = F\left(\nabla^2 u, \nabla u, u, x\right)$ be a second-order operator with F continuous in its Hessian argument at the origin. Assume its residual is square integrable for the networks considered, and let $\mathcal{F}_0[u](x) = F(0, \nabla u, u, x)$ be its reduced first-order part. In this theorem the risk is $J_{\mathcal{F}}(u) = \|\mathcal{F}[u] - f\|^2 + \mu B(u)$, with B as in Theorem 2.4. For every $u \in \mathcal{N}_{\mathsf{PL}}$ one has $\mathcal{F}[u] = \mathcal{F}_0[u]$ almost everywhere, so the interior residual is $\|\mathcal{F}_0[u] - f\|^2$. Two regimes occur.*

(i) No interior signal. *If $\mathcal{F}_0[u] = q \in L^2(\Omega)$ depends on neither ∇u nor u, in particular $q \equiv 0$ when F is homogeneous of positive degree in the Hessian, then the interior residual equals the constant $\|q - f\|^2$ for every network. When $f \neq q$ on a set of positive measure this is a strictly positive floor, the parameter gradient of the interior term vanishes identically, and training is governed by the boundary functional alone.*

(ii) Reduced first-order equation. *Suppose $F(0, p, z, x)$ is Lipschitz in (p, z) uniformly in x with $F(0, 0, 0, \cdot) \in L^2(\Omega)$, let $\mu = 0$, and suppose $\mathcal{F}_0[w] = f$ has a solution in $H^1(\Omega)$. Then $\inf_{u \in \mathcal{N}_{\mathsf{PL}}} J_{\mathcal{F}}(u) = 0$, and every H^1-limit point $\tilde{u}$ of a minimizing sequence for $J_{\mathcal{F}}$ solves $\mathcal{F}_0[\tilde{u}] = f$ almost everywhere. If a true solution $u^\star \in H^1(\Omega)$ is twice differentiable almost everywhere and $\mathcal{F}[u^\star] = f$ almost everywhere, then $\tilde{u} = u^\star$ only if $F\left(\nabla^2 u^\star, \nabla u^\star, u^\star, x\right) = F(0, \nabla u^\star, u^\star, x)$ almost everywhere.*

Examples of both regimes, Monge–Ampère and the p-Laplacian for the first and Hamilton–Jacobi–Bellman operators for the second, are worked out in Theorem B.2 in Section B.

4.2 The Degree–Order Ladder and Score Matching

Depth raises the attainable derivative order only when the degree of the activation exceeds one.

Proposition 4.3 (Depth, degree, and the wall location; restatement of Theorem 1.4; proved in Section B)**.** *Let u be a network of depth L with activations in P_p. Then there is a finite partition of Ω with Lebesgue-null interfaces into cells cut out by semialgebraic sets, by polytopes when $p = 1$, such that on the interior of each cell u is a polynomial of degree at most p^L. Hence $D^\alpha u = 0$ almost everywhere for every α with $|\alpha| > p^L$. So the wall of u sits at order at most $p^L + 1$, and the degree-p rectified power unit attains this bound. In particular,*

(i) *for $p = 1$ the wall is at order at most 2 for every depth L, recovering Theorem 2.6; and*

(ii) *for $p \geq 2$, a nonzero derivative of order r requires depth $L \geq \log_p r$.*

Any objective that differentiates a learned scalar potential inherits the collapse one order down; score matching is the case we make exact.

Definition 4.4 (Energy parametrization and the score-matching objective)**.** Let q be a probability density on Ω, continuously differentiable, with $\mathbb{E}_{x \sim q} |\nabla \log q(x)|^2 < \infty$. A model assigns to a scalar energy φ_θ the density $p_\theta \propto e^{\varphi_\theta}$ and the score $s_\theta = \nabla \log p_\theta = \nabla \varphi_\theta$. The Fisher divergence

and the score-matching objective are

$$D_F(q, p_\theta) = \frac{1}{2}\mathbb{E}_{x\sim q}\left[\left|\nabla \log q(x) - s_\theta(x)\right|^2\right], \qquad J_{\mathrm{SM}}(\theta) = \mathbb{E}_{x\sim q}\left[\frac{1}{2}\left|\nabla\varphi_\theta(x)\right|^2 + \Delta\varphi_\theta(x)\right].$$

Under the conditions of an identity of Hyvärinen, which include a differentiable score and are recalled as Theorem B.3 in Section B, the objective J_{SM} equals the Fisher divergence up to a constant. On a piecewise-linear energy, where the identity does not apply, the second-order term of J_{SM} vanishes almost everywhere, and the objective collapses.

Theorem 4.5 (Score-matching collapse for piecewise-linear energies; restatement of Theorem 1.6; proved in Section B)**.** *Let the energy be parametrized by $\varphi_\theta \in \mathcal{N}_{\mathsf{PL}}$, and let the objective be evaluated, as in training, with the almost-everywhere derivatives of Theorem 2.2. Then $\Delta\varphi_\theta = 0$ almost everywhere, the objective reduces to*

$$J_{\mathrm{SM}}(\theta) = \mathbb{E}_{x\sim q}\left[\frac{1}{2}\left|\nabla\varphi_\theta(x)\right|^2\right],$$

and its infimum over $\mathcal{N}_{\mathsf{PL}}$ is zero, attained by the constant energy, for which the recovered score is $s_\theta \equiv 0$. Whenever q is not the uniform density, the Fisher divergence at this minimizer equals $C = \frac{1}{2}\mathbb{E}_q[|\nabla \log q|^2] > 0$; hence the procedure is inconsistent: its objective vanishes at the trivial score, whatever the data.

The singular term that the training convention drops is described in Theorem B.4 in Section B.

5 The Collapse Under Optimization

The collapse survives the passage from the idealized risk to the objective training optimizes.

5.1 The Limits of Repair

A penalty evaluated at the almost-everywhere Hessian cannot restore the curvature that the strong-form residual loses. First-order penalties can change the limit, as the example in Theorem C.2 shows.

Theorem 5.1 (Regularization within the autodiff regime cannot read piecewise-linear curvature; restatement of Theorem 1.5; proved in Section C)**.** *Let $R[u] = \int_\Omega r\left(\nabla^2 u, \nabla u, u, x\right) dx$ be a regularizer evaluated at the almost-everywhere derivatives, finite on $\mathcal{N}_{\mathsf{PL}}$, and set $J_\lambda = J + \lambda R$ for $\lambda \geq 0$.*

(a) *For every $u \in \mathcal{N}_{\mathsf{PL}}$ one has $R[u] = \int_\Omega r\,(0, \nabla u, u, x)\,dx$; in particular $\int_\Omega \left|\nabla^2 u\right|^2 dx = 0$ for every $u \in \mathcal{N}_{\mathsf{PL}}$.*

(b) *If r depends on its Hessian argument and on x alone, then R is the constant $R_0 = \int_\Omega r\,(0, x)\,dx$ on $\mathcal{N}_{\mathsf{PL}}$. For every $\lambda \geq 0$, adding λR to J or to $J_{\mathcal{F}}$ preserves the corresponding minimizing sequences and limits of Theorems 3.2, 4.1 and 4.2, while shifting each infimum by λR_0.*

Theorem C.2 in Section C compares part (a) with the Sobolev penalty of Doumèche et al. (2025). The remark also shows, for the model problem and $\lambda = 1$, that the gradient penalty $\int_\Omega |\nabla u|^2\,dx$ alone moves the limit to the solution of a Robin problem, which for $u^\star \in H^2(\Omega)$ is $u^\star$ only when $\partial_\nu u^\star = 0$. Theorem C.1 there says why the conforming finite element method, a weak formulation, is consistent on continuous piecewise-linear functions for the model problem with $g = 0$ on a polyhedral domain, and why a trained weak form need not inherit that consistency.

5.2 Training Reaches the Spurious Limit

Collocation makes the reduction exact, and gradient descent reaches the spurious limit when it trains the output weights of fixed features.

Proposition 5.2 (Exactness under collocation; restatement of Theorem 1.7; proved in Section C)**.** *Let $x_1, \dots, x_n \in \Omega$ be collocation points, fix a representative of f, and let the model operator be evaluated at the points as in training (Theorem 2.2). Then for every $u \in \mathcal{N}_{\mathsf{PL}}$ the empirical interior risk satisfies*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{L}u(x_i) - f(x_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(u(x_i) - f(x_i)\right)^2,$$

the empirical regression misfit to f, into which the equation enters through no term; its infimum over $\mathcal{N}_{\mathsf{PL}}$ is zero, attained by every network interpolating f at the collocation points, and such networks exist. If the points are drawn from a law absolutely continuous with respect to Lebesgue measure, then for every fixed u they almost surely avoid the interfaces of u, and the identity holds with the classical derivatives as well.

With fixed features, the limit of training itself tends to the spurious limit as their spans become dense.

Theorem 5.3 (Gradient descent reaches the spurious limit with fixed features; restatement of Theorem 1.8; proved in Section C)**.** *Fix features $\phi_1, \dots, \phi_m \in \mathcal{N}_{\mathsf{PL}}$, such as a fixed piecewise-linear hidden layer, and train only the output weights $\theta \in \mathbb{R}^m$, so that $u_\theta = \sum_{j=1}^m \theta_j \phi_j$. For the model operator with $\mu = 0$ the physics-informed risk is the convex quadratic $J(\theta) = \|u_\theta - f\|^2 = \theta^\top \Gamma \theta - 2\theta^\top \beta + \|f\|^2$ with $\Gamma_{jk} = \langle \phi_j, \phi_k \rangle$ and $\beta_j = \langle \phi_j, f \rangle$. Gradient flow $\dot\theta = -\nabla J(\theta)$ converges from every initialization to a global minimizer $\theta^\star$, with $\Gamma\theta^\star = \beta$. If $\Gamma \succ 0$ the minimizer is unique and the convergence is at the linear rate $e^{-2\lambda_{\min}(\Gamma)t}$. Gradient descent has the same limit with a step size below $1/\lambda_{\max}(\Gamma)$ when $\Gamma \neq 0$, and with any step size when $\Gamma = 0$. The limit $u_{\theta^\star}$ is the L^2-orthogonal projection of f onto $\operatorname{span}\{\phi_j\}$. Along feature sets whose spans become dense in $L^2(\Omega)$, the limit tends to f, so its distance to the true solution tends to the inconsistency $\delta(f)$. For $\mu > 0$ the flow and descent statements hold with $\mu \langle \phi_j, \phi_k \rangle_{L^2(\partial\Omega)}$ added to Γ_{jk} and $\mu \langle \phi_j, g \rangle_{L^2(\partial\Omega)}$ added to β_j. The limit is then the minimizer of $\|u - f\|^2 + \mu \|u - g\|^2_{L^2(\partial\Omega)}$ over the span, and it tends to f along feature spans dense in $H^1(\Omega)$.*

6 Consistency of Smooth Activations

For a smooth activation, consistency follows from approximation alone.

Theorem 6.1 (Consistency of sigmoid solvers, with rate; restatement of Theorem 1.9; proved in Section D)**.** *Let $\sigma \in \mathsf{S}$, and let $s \geq 2$ be an integer. Let $u^\star$ solve $\mathcal{L}u^\star = f$ with $u^\star = g$ on $\partial\Omega$. Suppose $u^\star$ is the restriction to Ω of a function in $C^s(\mathbb{R}^d)$. Let $C_{\mathcal{L}}$ be such that $\|\mathcal{L}w\|_{L^\infty(\Omega)} \leq C_{\mathcal{L}} \|w\|_{W^{2,\infty}(\Omega)}$ for all $w \in W^{2,\infty}(\Omega)$.*

(i) $\inf_{u \in \mathcal{N}_{\mathsf{S}}} J(u) = 0$*; and if σ is the logistic sigmoid or the hyperbolic tangent and $s > 2$, then for every $\varepsilon > 0$ there are networks u with N nonzero weights and $J(u) = O\big(N^{-2(s-2-\varepsilon)/d}\big)$.*

(ii) *If $\mu > 0$ and $\|w\| \leq C_S \left(\|\mathcal{L}w\| + \|w\|_{L^2(\partial\Omega)}\right)$ for every $w \in C^2(\overline{\Omega})$, as holds for the model problem by Theorem D.1 in Section D, then every minimizing sequence converges to $u^\star$ in $L^2(\Omega)$; indeed $\|u - u^\star\| \leq C_S \left(1 + \mu^{-1/2}\right) J(u)^{1/2}$ on $\mathcal{N}_{\mathsf{S}}$, so the rate of* (i) *gives $\|u - u^\star\| = O\big(N^{-(s-2-\varepsilon)/d}\big)$.*

Part (ii) needs $\mu > 0$: the interior residual cannot see the nonzero solutions of $\mathcal{L}h = 0$. The approximation theory behind part (i) is traced in Section E.

AI use statement

The authors posed the problem. Claude (Anthropic) helped develop the proofs, locate relevant literature, and draft the manuscript. Codex (OpenAI) assisted with reviewing the mathematical arguments, citations, and prose of the ICLR submission. The authors reviewed the AI-assisted work and take responsibility for the paper and any errors.

Under the ICLR 2027 AI Policy for Authors, framework development, result formulation, proof ingredients, and proof writing require disclosure; drafting, literature summary, and reference formatting are recommended for disclosure. Experiments, data, research code, and translation do not arise in this paper.

Reproducibility Statement

Every result in this paper is analytical and there are no experiments, no data and no code. Each hypothesis is stated in the numbered definition it belongs to, in Section 2, and each result carries a complete proof: proofs are collected in Sections A to D, one appendix section per body section, with a pointer beside every statement. Each result imported from the literature is stated as a numbered fact with its citation and is never used beyond what that statement gives.

References

D. Belomestny, A. Naumov, N. Puchkin, and S. Samsonov. Simultaneous approximation of a smooth function and its derivatives by deep neural networks with piecewise-polynomial activations. *Neural Networks*, 161:242–253, 2023.

S. Berrone, C. Canuto, and M. Pintore. Variational physics informed neural networks: The role of quadratures and test functions. *Journal of Scientific Computing*, 92(3):100, 2022. doi: 10.1007/s10915-022-01950-4.

C. Chen, Y. Yang, Y. Xiang, and W. Hao. Automatic differentiation is essential in training neural networks for solving differential equations. *Journal of Scientific Computing*, 104(2):54, 2025.

X. Chen, P. Rosakis, Z. Wu, and Z. Zhang. Solving nonconvex energy minimization problems in martensitic phase transitions with a mesh-free deep learning approach. *Computer Methods in Applied Mechanics and Engineering*, 416:116384, 2023.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

T. De Ryck, S. Lanthaler, and S. Mishra. On the approximation of functions by $\tanh$ neural networks. *Neural Networks*, 143:732–750, 2021.

R. A. DeVore, R. Howard, and C. A. Micchelli. Optimal nonlinear approximation. *Manuscripta Mathematica*, 63(4):469–478, 1989. doi: 10.1007/BF01171759.

N. Doumèche, G. Biau, and C. Boyer. On the convergence of PINNs. *Bernoulli*, 31(3):2127–2151, 2025. doi: 10.3150/24-BEJ1799.

W. E and B. Yu. The deep Ritz method: A deep learning-based numerical algorithm for solving variational problems. *Communications in Mathematics and Statistics*, 6(1):1–12, 2018. doi: 10.1007/s40304-018-0127-z.

L. C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, 2nd edition, 2010.

I. Gühring and M. Raslan. Approximation rates for neural networks with encodable weights in smoothness spaces. *Neural Networks*, 134:107–130, 2021.

I. Gühring, G. Kutyniok, and P. Petersen. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Analysis and Applications*, 18(5):803–859, 2020. doi: 10.1142/S0219530519410021.

K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2): 251–257, 1991.

K. Hornik, M. Stinchcombe, H. White, and P. Auer. Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. *Neural Computation*, 6(6): 1262–1275, 1994.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

A. Jentzen and A. Riekert. On the existence of global minima and convergence analyses for gradient descent methods in the training of deep neural networks. *Journal of Machine Learning*, 1(2): 141–246, 2022. doi: 10.4208/jml.220114a.

Y. Jiao, Y. Lai, D. Li, X. Lu, F. Wang, Y. Wang, and J. Z. Yang. A rate of convergence of physics informed neural networks for the linear second order elliptic PDEs. *Communications in Computational Physics*, 31(4):1272–1295, 2022. doi: 10.4208/cicp.OA-2021-0186.

G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

I. E. Lagaris, A. Likas, and D. I. Fotiadis. Artificial neural networks for solving ordinary and partial differential equations. *IEEE Transactions on Neural Networks*, 9(5):987–1000, 1998.

W. Lee, H. Yu, X. Rival, and H. Yang. On correctness of automatic differentiation for non-differentiable functions. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6719–6730, 2020.

K. Leng and J. Thiyagalingam. On the compatibility between neural networks and partial differential equations for physics-informed learning, 2022. Preprint, arXiv:2212.00270.

M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a non-polynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

B. Li, S. Tang, and H. Yu. PowerNet: Efficient representations of polynomials and smooth functions by deep neural networks with rectified power units. *Journal of Mathematical Study*, 53(2):159–191, 2020. doi: 10.4208/jms.v53n2.20.03.

Y. Li, S. Lu, P. Mathé, and S. V. Pereverzev. Two-layer networks with the ReLU^k activation function: Barron spaces and derivative approximation. *Numerische Mathematik*, 156(1):319–344, 2024.

H. N. Mhaskar. Neural networks for optimal approximation of smooth and analytic functions. *Neural Computation*, 8(1):164–177, 1996.

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

S. Rojas, P. Maczuga, J. Muñoz-Matute, D. Pardo, and M. Paszyński. Robust variational physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 425: 116904, 2024. doi: 10.1016/j.cma.2024.116904.

Conor Rowan. On the failure of ReLU activation for physics-informed machine learning, 2025. URL `https://arxiv.org/abs/2512.11184`. Preprint, arXiv:2512.11184.

G. Shen, Y. Jiao, Y. Lin, and J. Huang. Differentiable neural networks with RePU activation: With applications to score estimation and isotonic regression, 2023. Preprint, arXiv:2305.00608.

Y. Shin, J. Darbon, and G. E. Karniadakis. On the convergence of physics informed neural networks for linear second-order elliptic and parabolic type PDEs. *Communications in Computational Physics*, 28(5):2042–2074, 2020. doi: 10.4208/cicp.OA-2020-0193.

C. Wang, S. Li, D. He, and L. Wang. Is L^2 physics informed loss always suitable for training physics informed neural network? In *Advances in Neural Information Processing Systems*, volume 35, pp. 8278–8290, 2022.

A Proofs for the Inconsistency Theorem

Proof of Theorem 2.8. The rectified linear unit is continuous and non-polynomial, so by the density criterion of Leshno et al. (1993) the class $\mathcal{N}_{\mathsf{PL}}$ is dense in $C(\overline{\Omega})$, and $C(\overline{\Omega})$ is dense in $L^2(\Omega)$. For the H^1 density let $w \in H^1(\Omega)$ and $\delta > 0$. Since $\partial\Omega$ is of class C^1, the function w extends to a function $Ew \in H^1(\mathbb{R}^d)$ with compact support (Evans, 2010, Section 5.4, Theorem 1). The mollifications of Ew converge to it in H^1 on every bounded set (Evans, 2010, Section 5.3, Theorem 1). Fix a cube $Q \supset \overline{\Omega}$ and a mollification $w_\delta \in C_c^\infty(\mathbb{R}^d)$ with $\|Ew - w_\delta\|_{H^1(Q)} \le \delta$. As $w_\delta \in W^{2,2}(Q)$, we may apply the Sobolev approximation theorem of Gühring et al. (2020) for the rectified linear unit, which is stated on $(0,1)^d$ and carries to Q by an affine change of variables. Applied with regularity 2, integrability 2 and target smoothness 1, it gives a rectified-linear network $\Phi \in \mathcal{N}_{\mathsf{PL}}$ with $\|w_\delta - \Phi\|_{W^{1,2}(Q)} \le \delta$. Hence $\|w - \Phi\|_{H^1(\Omega)} \le \|Ew - w_\delta\|_{H^1(Q)} + \|w_\delta - \Phi\|_{H^1(Q)} \le 2\delta$. □

Proof of Theorem 3.1. Writing the principal part in non-divergence form,

$$-\operatorname{div}(A\nabla u) = -\sum_{i,j} \partial_i \left(A_{ij}\partial_j u\right) = -\sum_{i,j} \left(\partial_i A_{ij}\right)\partial_j u - \sum_{i,j} A_{ij}\partial_i\partial_j u,$$

where A is Lipschitz, so $\partial_i A_{ij}$ exists almost everywhere and is bounded. The last sum, with A_{ij} the entries of A, is the contraction of A with the Hessian of u, which is zero almost everywhere by Theorem 2.6. The first sum is $-(\nabla\cdot A)\cdot\nabla u$. Collecting the surviving terms with $b\cdot\nabla u + cu$ gives the stated first-order operator. For $\mathcal{L} = -\Delta + 1$ one has $A = I$, $b \equiv 0$, $c = 1$, so the gradient terms vanish as well and $\mathcal{L}u = u$ almost everywhere. □

Proof of Theorem 3.2. Part (a). By Theorem 3.1 with the model operator, $\mathcal{L}u = u$ almost everywhere, so $(\mathcal{L}u - f)^2 = (u - f)^2$ almost everywhere; integrating over Ω and adding $\mu B(u) \ge 0$ gives the claim.

Part (b). Let $w_g \in H^1(\Omega)$ be a function with trace g. Choose $f_n \in C_c^\infty(\Omega)$ with $f_n \to f - w_g$ in $L^2(\Omega)$, and set $h_n = f_n + w_g \in H^1(\Omega)$, whose trace is g. By the H^1 density of Theorem 2.8 choose $u_n \in \mathcal{N}_{\mathsf{PL}}$ with $\|u_n - h_n\|_{H^1(\Omega)} \le 1/n$. Then $\|u_n - f\| \le \|u_n - h_n\| + \|f_n - (f - w_g)\| \to 0$, so by part (a) the interior residual of u_n tends to zero, while the trace theorem (Evans, 2010, Section 5.5, Theorem 1), with its constant C depending only on Ω, gives $B(u_n) = \|u_n - g\|^2_{L^2(\partial\Omega)} = \|u_n - h_n\|^2_{L^2(\partial\Omega)} \le C^2\|u_n - h_n\|^2_{H^1(\Omega)} \to 0$. Hence $J(u_n) \to 0$, and by part (a) $J \ge 0$, so the infimum is zero.

Part (c). Let (u_n) be any minimizing sequence. By part (a), $\int_\Omega (u_n - f)^2\,dx \le J(u_n) \to 0$, so $u_n \to f$ in $L^2(\Omega)$ and therefore $\|u_n - u^\star\| \to \|f - u^\star\|$. From Eq. (2), $f - u^\star = (-\Delta u^\star + u^\star) - u^\star = -\Delta u^\star$ in the weak sense, so $\|f - u^\star\| = \|\Delta u^\star\| = \delta(f)$. The training residual tends to zero by the choice of a minimizing sequence. If $u^\star$ is non-harmonic then $\Delta u^\star \ne 0$ on a set of positive measure, so $\delta(f) > 0$.

Part (d). Let (φ_k) be an $L^2(\Omega)$-orthonormal basis of eigenfunctions of $-\Delta$ with zero Dirichlet data, with eigenvalues $\lambda_k \to \infty$ (Evans, 2010, Section 6.5, Theorem 1). Take the source $f = \varphi_k$ with data $g = 0$. Then $u^\star = \varphi_k/(1+\lambda_k)$, so $\|f\| = 1$, the error is $\delta(f) = \|f - u^\star\| = \lambda_k/(1+\lambda_k)$, and $\|u^\star\| = 1/(1+\lambda_k)$. A k with $\lambda_k \ge 1/\varepsilon - 1$ gives $\delta(f) \ge 1 - \varepsilon$ and $\|u^\star\| \le \varepsilon$, and $\delta(f)/\|u^\star\| = \lambda_k$ is unbounded. □

On the torus, where the Fourier basis diagonalizes the model operator, a family like that of Theorem 3.2(d) is explicit.

Proposition A.1 (Spectral quantification on the torus)**.** *Let $\Omega = \mathbb{T}^d$ be the flat torus and $\mathcal{L} = -\Delta + 1$. For f with Fourier coefficients $\hat{f}(\boldsymbol{\xi})$ indexed by $\boldsymbol{\xi} \in \mathbb{Z}^d$ and normalized so that $\|f\|^2 = \sum_{\boldsymbol{\xi}} \left|\hat{f}(\boldsymbol{\xi})\right|^2$, the solution is $\hat{u}^\star(\boldsymbol{\xi}) = \hat{f}(\boldsymbol{\xi})/\left(1 + |\boldsymbol{\xi}|^2\right)$ and*

$$\delta(f)^2 = \|f - u^\star\|^2 = \sum_{\boldsymbol{\xi}\in\mathbb{Z}^d} \left|\hat{f}(\boldsymbol{\xi})\right|^2 \left(\frac{|\boldsymbol{\xi}|^2}{1 + |\boldsymbol{\xi}|^2}\right)^2 = \|\Delta u^\star\|^2. \tag{3}$$

If f is supported on frequencies $|\boldsymbol{\xi}| \geq K$ with $\|f\| = 1$, then

$$\frac{\|f - u^\star\|^2}{\|f\|^2} \geq \left(\frac{K^2}{1+K^2}\right)^2 \xrightarrow[K\to\infty]{} 1, \qquad \|u^\star\|^2 \leq \frac{1}{(1+K^2)^2} \xrightarrow[K\to\infty]{} 0.$$

In particular such an f has $\delta(f) \geq K^2/\left(1+K^2\right)$ and $\|u^\star\| \leq 1/\left(1+K^2\right)$, so $\delta(f)/\|u^\star\| \geq K^2$. As K grows, the inconsistency $\delta(f)$ tends to the full norm $\|f\| = 1$ while $\|u^\star\|$ tends to zero, as in Theorem 3.2(d)*, here in closed form.*

Proof of Theorem A.1. On $\mathbb{T}^d$ the operator is diagonalized by the Fourier basis $e^{i\boldsymbol{\xi}\cdot x}$ with symbol $1 + |\boldsymbol{\xi}|^2$, giving $\hat{u}^\star(\boldsymbol{\xi}) = \hat{f}(\boldsymbol{\xi})/\left(1+|\boldsymbol{\xi}|^2\right)$. Then $\widehat{f - u^\star}(\boldsymbol{\xi}) = \hat{f}(\boldsymbol{\xi})\,|\boldsymbol{\xi}|^2/\left(1+|\boldsymbol{\xi}|^2\right)$ and $\widehat{\Delta u^\star}(\boldsymbol{\xi}) = -|\boldsymbol{\xi}|^2\,\hat{u}^\star(\boldsymbol{\xi}) = -\hat{f}(\boldsymbol{\xi})\,|\boldsymbol{\xi}|^2/\left(1+|\boldsymbol{\xi}|^2\right)$, so the two Fourier representations agree up to sign, hence have equal norms, and Parseval gives Eq. (3). The map $t \mapsto t/(1+t)$ is increasing, so on frequencies $|\boldsymbol{\xi}|^2 \geq K^2$ each factor $|\boldsymbol{\xi}|^2/\left(1+|\boldsymbol{\xi}|^2\right)$ is at least $K^2/\left(1+K^2\right)$, which gives the relative-error bound. Summing $|\hat{u}^\star(\boldsymbol{\xi})|^2 \leq \left|\hat{f}(\boldsymbol{\xi})\right|^2/\left(1+K^2\right)^2$ over the support gives the bound on $\|u^\star\|^2$. Each factor $|\boldsymbol{\xi}|^2/\left(1+|\boldsymbol{\xi}|^2\right)$ is below one, so Eq. (3) also gives $\delta(f) \leq \|f\|$, and letting $K \to \infty$ proves the limits. □

B Proofs for the Generality and Transfer of the Collapse

Proof of Theorem 4.1. By Theorem 3.1, $\mathcal{L}u = \tilde{\mathcal{L}}u$ almost everywhere for $u \in \mathcal{N}_{\mathsf{PL}}$, so the interior residual equals $\left\|\tilde{\mathcal{L}}u - f\right\|^2$. Since $b \cdot \nabla u$ and $(\nabla \cdot A) \cdot \nabla u$ are the only first-order terms of $\mathcal{L}u$, the condition $\tilde{b} \equiv 0$ holds exactly when $\mathcal{L}u = -\sum_{i,j} A_{ij}\partial_i\partial_j u + cu$.

Part (i). With $\tilde{b} \equiv 0$ the interior residual is $\|cu - f\|^2 = \|c\,(u - f/c)\|^2$, which lies between $c_0^2\,\|u - f/c\|^2$ and $\|c\|_{L^\infty}^2\,\|u - f/c\|^2$ because $c_0 \leq c \leq \|c\|_{L^\infty}$. As $f/c \in L^2(\Omega)$, we may run the proof of Theorem 3.2(b) with f/c in place of f. It produces networks $u_n \in \mathcal{N}_{\mathsf{PL}}$ with $\|u_n - f/c\| \to 0$ and $B(u_n) \to 0$. Hence $J(u_n) \to 0$, and the infimum is zero. For a minimizing sequence, $c_0^2\,\|u_n - f/c\|^2 \leq J(u_n) \to 0$, so $u_n \to f/c$ in $L^2(\Omega)$. Finally $f/c = u^\star$ holds exactly when $f = cu^\star$, and since $f = \mathcal{L}u^\star = -\sum_{i,j} A_{ij}\partial_i\partial_j u^\star + cu^\star$ in the sense of distributions, this is the vanishing of the second-order part.

Part (ii). Let $w \in H^1(\Omega)$ solve $\tilde{\mathcal{L}}w = f$ and, by Theorem 2.8, let $u_n \in \mathcal{N}_{\mathsf{PL}}$ converge to w in $H^1(\Omega)$. Since $\tilde{b}$ and c are bounded, $\tilde{\mathcal{L}}u_n \to \tilde{\mathcal{L}}w = f$ in $L^2(\Omega)$, so $J(u_n) = \left\|\tilde{\mathcal{L}}u_n - f\right\|^2 \to 0$ and the infimum is zero. Let (u_n) be a minimizing sequence with a subsequence, not relabelled, converging to $\tilde{u}$ in $L^2(\Omega)$; then $\tilde{\mathcal{L}}u_n \to f$ in $L^2(\Omega)$. Fix a test function $\psi \in C_c^\infty(\Omega)$. Each u_n is Lipschitz, and $\tilde{b}\psi$ is Lipschitz with compact support, so integration by parts gives

$$\int_\Omega \left(\tilde{b} \cdot \nabla u_n\right)\psi\,dx = -\int_\Omega u_n \operatorname{div}\left(\tilde{b}\psi\right)dx \to -\int_\Omega \tilde{u} \operatorname{div}\left(\tilde{b}\psi\right)dx,$$

while $\int_\Omega cu_n\psi\,dx \to \int_\Omega c\tilde{u}\psi\,dx$. Since $\tilde{\mathcal{L}}u_n \to f$ in $L^2(\Omega)$, the two limits add up to

$$\int_\Omega f\psi\,dx = -\int_\Omega \tilde{u} \operatorname{div}\left(\tilde{b}\psi\right)dx + \int_\Omega c\tilde{u}\psi\,dx,$$

which is the equation $\tilde{\mathcal{L}}\tilde{u} = f$ in the sense of distributions. If $\tilde{u} = u^\star$, then $\tilde{\mathcal{L}}u^\star = f = \mathcal{L}u^\star$ as distributions. The difference $\mathcal{L}u^\star - \tilde{\mathcal{L}}u^\star = -\operatorname{div}(A\nabla u^\star) + (\nabla \cdot A) \cdot \nabla u^\star$ is the second-order part of $\mathcal{L}$ applied to $u^\star$, which is $-\sum_{i,j} A_{ij}\partial_i\partial_j u^\star$ when $u^\star \in H^2(\Omega)$. □

Part (ii) is stated for the interior risk, and the restriction is needed.

Remark B.1 (The drift case needs the interior risk)**.** Part (ii) cannot be stated for the full risk. Take $\Omega = (0,1)$ and $\mathcal{L}u = -u'' + u' + u$, so that $A = 1$, $b = c = 1$ and the reduced operator is $\tilde{\mathcal{L}}u = u' + u$. Take the source $f = 0$ and the data $g(0) = 0$ and $g(1) = 1$, which are the trace of $w_g(x) = x$; the boundary term is then $B(u) = |u(0)|^2 + |u(1) - 1|^2$. For every $\mu > 0$ the infimum of J over $\mathcal{N}_{\mathsf{PL}}$ is positive. Indeed, suppose networks $u_n \in \mathcal{N}_{\mathsf{PL}}$ had $J(u_n) \to 0$. Then $u_n' + u_n \to 0$ in $L^2(0,1)$ by Theorem 3.1, and the boundary term forces $u_n(0) \to 0$ and $u_n(1) \to 1$. The function $v_n = e^x u_n$ is Lipschitz with $v_n' = e^x (u_n' + u_n)$ almost everywhere, so by the Cauchy–Schwarz inequality $|v_n(1) - v_n(0)| \le e \|u_n' + u_n\| \to 0$, whereas $v_n(1) - v_n(0) = e u_n(1) - u_n(0) \to e$. The reduced equation $w' + w = 0$ is first order, and its solutions $w(0)e^{-x}$ take only one boundary value freely, so the two boundary conditions and the reduced residual cannot be met together. The interior risk alone has infimum zero, by part (ii) with $w = 0$.

Proof of Theorem 4.2. A piecewise-linear u is affine on each cell of its partition, so $\nabla^2 u = 0$ almost everywhere, and continuity of F at the origin in the Hessian slot gives $\mathcal{F}[u] = F(0, \nabla u, u, x) = \mathcal{F}_0[u]$ almost everywhere; the interior residual is thus $\|\mathcal{F}_0[u] - f\|^2$. In case (i) this equals $\|q - f\|^2$, independent of the parameters, so its parameter gradient is zero and $J_{\mathcal{F}}(u) = \|q - f\|^2 + \mu B(u)$ is minimized over the boundary term alone.

In case (ii), let Λ be the Lipschitz constant of $F(0, \cdot, \cdot, x)$. For $v, w \in H^1(\Omega)$ one has $|\mathcal{F}_0[v] - \mathcal{F}_0[w]| \le \Lambda(|\nabla v - \nabla w| + |v - w|)$ pointwise, so $\|\mathcal{F}_0[v] - \mathcal{F}_0[w]\| \le \sqrt{2}\Lambda \|v - w\|_{H^1(\Omega)}$, and $\mathcal{F}_0[v] \in L^2(\Omega)$ because $|\mathcal{F}_0[v]| \le |F(0,0,0,x)| + \Lambda(|\nabla v| + |v|)$. Let $w \in H^1(\Omega)$ solve $\mathcal{F}_0[w] = f$ and, by Theorem 2.8, let $u_n \in \mathcal{N}_{\mathsf{PL}}$ converge to w in $H^1(\Omega)$. Then $J_{\mathcal{F}}(u_n) = \|\mathcal{F}_0[u_n] - \mathcal{F}_0[w]\|^2 \le 2\Lambda^2 \|u_n - w\|^2_{H^1(\Omega)} \to 0$, so the infimum is zero. Suppose a minimizing sequence has a subsequence converging in $H^1(\Omega)$ to $\tilde{u}$. The same bound gives $\mathcal{F}_0[u_n] \to \mathcal{F}_0[\tilde{u}]$ in $L^2(\Omega)$ along it, and $\mathcal{F}_0[u_n] \to f$ as well, so $\mathcal{F}_0[\tilde{u}] = f$ almost everywhere. Finally $\tilde{u} = u^\star$ forces $\mathcal{F}_0[u^\star] = f = \mathcal{F}[u^\star]$, that is $F\left(\nabla^2 u^\star, \nabla u^\star, u^\star, x\right) = F(0, \nabla u^\star, u^\star, x)$ almost everywhere. □

Three standard nonlinear operators illustrate the two regimes of Theorem 4.2.

Remark B.2 (Monge–Ampère, the p-Laplacian, and viscosity selection)**.** The Monge–Ampère operator $\det \nabla^2 u$ and the p-Laplacian $\operatorname{div}\left(|\nabla u|^{p-2} \nabla u\right)$ for $p \ge 2$ fall under case (i) of Theorem 4.2: a piecewise-linear network has piecewise-constant gradient, so each is zero almost everywhere and training sees the boundary data alone. Case (ii) of Theorem 4.2 leaves a genuine first-order equation. For a Hamilton–Jacobi–Bellman operator $\sup_a \left[-\operatorname{tr}\left(A_a \nabla^2 u\right) + b_a \cdot \nabla u + c_a u\right]$ with control-indexed coefficients, the reduced equation is $\sup_a [b_a \cdot \nabla u + c_a u] = f$. It can have many almost-everywhere solutions. The viscosity criterion (Evans, 2010), which selects among them, is invisible to residual minimization.

Proof of Theorem 4.3. We argue by induction on the layer ℓ. There is a finite partition of Ω into cells cut out by semialgebraic sets. On the interior of each cell, every coordinate of the output of layer ℓ is a polynomial of degree at most p^ℓ. The input (layer 0) is affine, of degree $p^0 = 1$, on the single cell Ω. Suppose the claim holds for layer $\ell - 1$, and fix one of its cells.

On that cell the preactivation $A_{\ell-1}(\cdot) + b_{\ell-1}$ is a vector of polynomials of degree at most $p^{\ell-1}$, the affine map preserving degree. The activation $\sigma \in \mathsf{P}_p$ is a single polynomial branch on each of finitely many intervals. For each coordinate of the preactivation and each such interval, the set of x in the cell at which that coordinate lies in that interval is cut out of the cell by polynomial inequalities. Intersecting over the coordinates refines the cell into finitely many subcells, each cut out of Ω by semialgebraic sets. On each subcell, σ applied entrywise composes a polynomial of degree at most p with one of degree at most $p^{\ell-1}$, which produces degree at most $p \cdot p^{\ell-1} = p^\ell$. Every interface of the refined partition lies in the zero set of one of the finitely many nonzero polynomials defining the branch regions, a Lebesgue-null set; when $p = 1$ those polynomials are affine and the cells are cut out by polytopes. Hence after L layers and a final affine map the output is cellwise of degree at most p^L. Any derivative of order exceeding p^L of a polynomial of degree at most p^L is zero, which gives the wall at order $p^L + 1$.

Setting $p = 1$ gives degree 1 at every depth, hence the order-2 wall and Theorem 2.6. For $p \geq 2$ a derivative of order r can survive only once $p^L \geq r$, that is, once $L \geq \log_p r$.

The bound is attained, so the wall of the class sits exactly at order $p^L + 1$. Take the degree-p rectified power unit $\sigma(t) = \max\{t, 0\}^p$ and any a with $\Omega \cap \{x_1 > a\}$ nonempty. The width-one network $\sigma(\cdots \sigma(x_1 - a) \cdots)$ with L activations equals $(x_1 - a)^{p^L}$ on that open set, since each activation raises a positive number to the power p. There its derivative $\partial_1^{p^L}$ is the nonzero constant $(p^L)!$; for $p = 1$ the network is $x_1 - a$, whose first derivative is 1. □

The Laplacian term of J_{SM} comes from an integration by parts due to Hyvärinen.

Fact B.3 (Hyvärinen's identity; Hyvärinen, 2005)**.** *Suppose q is positive in Ω, and q and the model score s_θ extend to $C^1(\overline{\Omega})$. Suppose also that $\mathbb{E}_q[|\nabla \log q|^2]$ is finite and that $q s_\theta \cdot \nu = 0$ on $\partial\Omega$, where ν is the outward unit normal. Then $D_F(q, p_\theta) = J_{\mathrm{SM}}(\theta) + C$, where $C = \frac{1}{2}\mathbb{E}_q[|\nabla \log q|^2]$ does not depend on θ (Hyvärinen, 2005, Theorem 1, stated on $\mathbb{R}^d$ with decay at infinity; the same integration by parts gives the bounded-domain identity). The term $\Delta\varphi_\theta$ of J_{SM} arises from integrating the cross term of $D_F(q, p_\theta)$ by parts.*

Proof of Theorem 4.5. Since $\varphi_\theta \in \mathcal{N}_{\mathsf{PL}}$ is continuous piecewise linear, its Hessian vanishes almost everywhere by Theorem 2.6, so its Laplacian $\Delta\varphi_\theta = \operatorname{tr} \nabla^2 \varphi_\theta$ is zero almost everywhere, which is the value used in training by Theorem 2.2. The law with density q is absolutely continuous, so the interfaces carry no mass. Substituting $\Delta\varphi_\theta = 0$ into J_{SM} leaves $\mathbb{E}_q[\frac{1}{2}|\nabla\varphi_\theta|^2] \geq 0$. The constant energy lies in $\mathcal{N}_{\mathsf{PL}}$ and gives $\nabla\varphi_\theta \equiv 0$, hence objective value 0; so the infimum is zero and is attained with $s_\theta = \nabla\varphi_\theta \equiv 0$. The Fisher divergence is defined through the score alone, so at $s_\theta \equiv 0$ it equals $\frac{1}{2}\mathbb{E}_q[|\nabla \log q|^2] = C$ by definition, with no identity invoked. Theorem B.3 does not carry the minimization of J_{SM} over to the Fisher divergence. It asks for a differentiable model score, and the gradient of an energy with a kink is not differentiable across the kink.

Suppose $C = 0$. Then $\nabla q = 0$ on the open set $\{q > 0\}$, so q is constant, hence positive, on each connected component U of that set. By continuity q is positive on the closure of U in Ω. So that closure lies inside $\{q > 0\}$, and, being connected and containing U, inside the component U; hence U is closed in Ω. As Ω is connected, $U = \Omega$, and q is the uniform density. Hence $C > 0$ whenever q is not uniform. □

What the training convention drops here is a measure on the kinks.

Remark B.4 (The discarded singular term)**.** The second-order term, here the divergence of the score, is computed as its almost-everywhere value. The distributional Laplacian of a piecewise-linear energy with at least one kink is a nonzero measure on the kinks, the measure that Hyvärinen's integration by parts requires and automatic differentiation discards. Indeed, integrating by parts cell by cell shows that the Laplacian of a continuous piecewise-linear φ acts on a test function ψ as $\sum_F c_F \int_F \psi \, dS$. Here F runs over the facets of the partition and c_F is the jump of the normal derivative of φ across F. Distinct facets carry mutually singular surface measures, so this measure is nonzero exactly when some $c_F \neq 0$, that is, when φ has a kink.

C Proofs for the Collapse Under Optimization

We first record why the collapse belongs to the strong form rather than to the class.

Remark C.1 (The collapse is a strong-form phenomenon)**.** The collapse comes from the strong, pointwise residual, through the reduction of Theorem 3.1, a pointwise computation that holds on any domain; in the following setting, a weak formulation over continuous piecewise-linear finite element functions converges to the solution instead. For the model problem with $g = 0$ on a polyhedral domain, a setting outside Theorem 2.1, continuous piecewise-linear functions are consistent under the weak formulation. There, the conforming P_1 finite element method seeks u among the continuous piecewise-linear functions on a triangulation that vanish on $\partial\Omega$, with $\int_\Omega \nabla u \cdot \nabla\varphi + \int_\Omega u\varphi = \int_\Omega f\varphi$ for every test φ among them. It converges as the mesh size of a shape-regular family of triangulations tends to zero, by coercivity and the density of these spaces in $H_0^1(\Omega)$. Integration by parts moves

one derivative onto the test function (Evans, 2010), so the weak residual reads the distributional Laplacian, including its part on the kinks, which the pointwise residual discards.

When the weak form is trained rather than assembled, with a network as trial class and a different test class, zero-residual spurious minimizers can reappear unless that pairing is inf-sup stable (Berrone et al., 2022). Whether the consistency of the finite element method above survives for trained piecewise-linear networks with an inf-sup-stable pairing is open. For the pointwise residual, no penalty on the almost-everywhere derivatives of a piecewise-linear network reads its curvature (Theorem 5.1), and a smooth saturating activation is a repair, which Theorem 6.1 proves consistent under its hypotheses.

Proof of Theorem 5.2. Evaluated as in training, the Hessian of u is zero at every point. Off the interfaces this is Theorem 2.6; on them, automatic differentiation implements the derivative of each activation as a piecewise-constant function whose own derivative it takes to be zero, so by the chain rule the Hessian it returns is again zero. So the computation of Theorem 3.1 gives $\mathcal{L}u(x_i) = u(x_i)$ at each collocation point, and substituting termwise gives the identity. The right side is nonnegative and vanishes exactly when $u(x_i) = f(x_i)$ for every i. Such a network exists. Let $y_1, \dots, y_k$ be the distinct points among the collocation points, let $h > 0$ be smaller than half the least ℓ^1-distance between two of them when $k \geq 2$, and let $\phi_i(x) = \max\{0, 1 - \|x - y_i\|_{\ell^1}/h\}$. Each ϕ_i is continuous and piecewise linear, and realizable by a rectified-linear network since $|t| = \rho(t) + \rho(-t)$; and $\phi_i(y_j) = \delta_{ij}$. So $u = \sum_{i=1}^{k} f(y_i)\phi_i$ lies in $\mathcal{N}_{\mathsf{PL}}$ and interpolates f at every collocation point. Finally, the interfaces of a fixed u lie in finitely many hyperplanes, a Lebesgue-null set, which points drawn from an absolutely continuous law avoid almost surely. At such a point u is affine on a neighborhood, so its classical Hessian is zero, and the identity is a statement about classical derivatives. □

Proof of Theorem 5.1. A piecewise-linear u has $\nabla^2 u = 0$ almost everywhere, so $r\left(\nabla^2 u, \nabla u, u, x\right) = r\left(0, \nabla u, u, x\right)$ almost everywhere, which is (a), and the curvature penalty integrates $|0|^2 = 0$. For (b), the same substitution gives $R[u] = \int_\Omega r(0, x)\, dx =: R_0$ for every $u \in \mathcal{N}_{\mathsf{PL}}$. Adding this constant to either J or $J_{\mathcal{F}}$ changes neither its minimizing sequences nor their limits and shifts its infimum by λR_0. Thus the claims of Theorems 3.2 and 4.1 transfer to $J_\lambda = J + \lambda R$, and those of Theorem 4.2 transfer to $J_{\mathcal{F}} + \lambda R$. □

Part (a) of Theorem 5.1 bears on a penalty from the literature, and it leaves room for penalties on first-order data, which can move the limit.

Remark C.2 (What a penalty can reach)**.** The Sobolev penalty $\lambda \|u\|^2_{H^{m+1}(\Omega)}$ of Doumèche et al. (2025, Definition 5.4) contains, for $m = 1$, the term $\lambda \int_\Omega \left|\nabla^2 u\right|^2 dx$ of Theorem 5.1(a), which is zero on $\mathcal{N}_{\mathsf{PL}}$. Doumèche et al. (2025, Section 2) also record, for the physics-informed penalty itself, that the high derivatives of a rectified-linear network vanish and so cannot be captured by it; Theorem 3.2 is the precise form of that observation. By Theorem 3.1 and part (a), the regularized risk is, on $\mathcal{N}_{\mathsf{PL}}$, a functional of ∇u and u alone, so no term of it reads the second-order content that distinguishes $u^\star$ from the solutions of the reduced equation.

A penalty on first-order data can still move the limit. One whose integrand depends on x can even encode $u^\star$. For the model problem and $\lambda > 0$, the integrand $r = (u - u^\star)^2 - (u - f)^2/\lambda$ makes $J_\lambda = \lambda \|u - u^\star\|^2 + \mu B$ on $\mathcal{N}_{\mathsf{PL}}$, but only by supplying the solution from outside the equation. The following example moves the limit with a penalty built from the equation alone. Take the model problem, $R[u] = \int_\Omega |\nabla u|^2 dx$ and $\lambda = 1$. On $\mathcal{N}_{\mathsf{PL}}$ the interior residual is $\|u - f\|^2$ by Theorem 3.1, and expanding the square gives $J_1(u) = \mathcal{Q}(u) + \|f\|^2$, where $\mathcal{Q}(u) = \|\nabla u\|^2 + \|u\|^2 - 2\langle f, u\rangle + \mu \|u - g\|^2_{L^2(\partial\Omega)}$ is defined on all of $H^1(\Omega)$. Write $a(u, v) = \langle \nabla u, \nabla v\rangle + \langle u, v\rangle + \mu \langle u, v\rangle_{L^2(\partial\Omega)}$ and $\ell(v) = \langle f, v\rangle + \mu \langle g, v\rangle_{L^2(\partial\Omega)}$, so that $\mathcal{Q}(u) = a(u, u) - 2\ell(u) + \mu \|g\|^2_{L^2(\partial\Omega)}$. The trace theorem (Evans, 2010, Section 5.5, Theorem 1) makes a and ℓ bounded on $H^1(\Omega)$. Moreover, $a(u, u) \geq \|u\|^2_{H^1(\Omega)}$. So the Lax–Milgram theorem (Evans, 2010, Section 6.2, Theorem 1) gives a unique $w \in H^1(\Omega)$ with $a(w, v) = \ell(v)$ for every $v \in H^1(\Omega)$. This w is the weak solution of $-\Delta w + w = f$ with the Robin condition $\partial_\nu w + \mu(w - g) = 0$ on $\partial\Omega$. For every $u \in H^1(\Omega)$, expand $a(u, u) = a(u - w, u - w) + 2a(w, u) - a(w, w)$. Using $a(w, u - w) = \ell(u - w)$ then gives

$\mathcal{Q}(u) - \mathcal{Q}(w) = a(u-w, u-w) \geq \|u-w\|^2_{H^1(\Omega)}$. Continuity and density (Theorem 2.8) show that the network infimum equals $\mathcal{Q}(w)$. By the identity just proved, every minimizing sequence of J_1 over $\mathcal{N}_{\mathsf{PL}}$ then converges to w in $H^1(\Omega)$.

So the limit is the Robin solution w, and for $u^\star \in H^2(\Omega)$ it is $u^\star$ only when $\partial_\nu u^\star = 0$ on $\partial\Omega$, a condition the equation does not impose. Indeed, suppose $u^\star \in H^2(\Omega)$ and $w = u^\star$. The strong equation holds in Ω, and $u^\star = g$ on $\partial\Omega$. Green's first identity therefore turns $a(u^\star, v) = \ell(v)$ into $\int_{\partial\Omega} v\partial_\nu u^\star \, dS = 0$ for every $v \in H^1(\Omega)$. Traces of $H^1(\Omega)$ functions include every function in $C^1(\partial\Omega)$, which is dense in $L^2(\partial\Omega)$, so $\partial_\nu u^\star = 0$ on $\partial\Omega$. What the pointwise residual lacks is the distributional Hessian, the measure on the kinks that automatic differentiation discards and that a smooth activation replaces with a classical Hessian.

Proof of Theorem 5.3. We compute the flow in closed form. Each u_θ is a linear combination of piecewise-linear functions, hence continuous piecewise linear, so by Theorem 3.1 the interior residual equals $\|u_\theta - f\|^2$, which expands into the displayed quadratic with the Gram matrix $\Gamma \succeq 0$ and the vector β. Its gradient is $\nabla J(\theta) = 2(\Gamma\theta - \beta)$. The normal equations $\Gamma\theta^\star = \beta$ are solvable, since β lies in the range of Γ. Indeed, let $Pf = \sum_k \theta^\star_k \phi_k$ be the L^2-orthogonal projection of f onto $V = \operatorname{span}\{\phi_j\}$; then $\beta_j = \langle \phi_j, f\rangle = \langle \phi_j, Pf\rangle = (\Gamma\theta^\star)_j$.

For any solution $\theta^\star$, gradient flow satisfies $\dot\theta = -2\Gamma(\theta - \theta^\star)$, so $\theta(t) - \theta^\star = e^{-2\Gamma t}(\theta(0) - \theta^\star)$, which converges to the projection of $\theta(0) - \theta^\star$ onto the kernel of Γ. The limit θ_∞ therefore satisfies $\Gamma\theta_\infty = \beta$ and is a global minimizer. If $\Gamma \succ 0$ the kernel is trivial, the minimizer is unique, and $|\theta(t) - \theta^\star| \leq e^{-2\lambda_{\min}(\Gamma)t}|\theta(0) - \theta^\star|$. Every minimizer solves the least-squares problem $\min_\theta \|u_\theta - f\|$, so $u_{\theta_\infty} = Pf$. Along feature spans V_n with $\operatorname{dist}(f, V_n) \to 0$ the projection error $\|f - Pf\|$ tends to zero, hence $u_{\theta_\infty} \to f$. By the triangle inequality and $\|f - u^\star\| = \delta(f)$, it follows that $\|u_{\theta_\infty} - u^\star\| \to \delta(f)$.

Gradient descent with step τ satisfies $\theta_{k+1} - \theta^\star = (I - 2\tau\Gamma)(\theta_k - \theta^\star)$. If $\Gamma \neq 0$ and $\tau < 1/\lambda_{\max}(\Gamma)$, every eigenvalue of $I - 2\tau\Gamma$ on the range of Γ lies in $(-1, 1)$, while the kernel is fixed. The iterates therefore converge to a minimizer with the same limit Pf. If $\Gamma = 0$, then every feature vanishes in $L^2(\Omega)$, $\beta = 0$, and every parameter vector is already a minimizer, for any step size.

For $\mu > 0$ the boundary term $B(u_\theta) = \left\|\sum_j \theta_j\phi_j - g\right\|^2_{L^2(\partial\Omega)}$ is a quadratic in θ. Its expansion adds $\mu\langle\phi_j, \phi_k\rangle_{L^2(\partial\Omega)}$ to Γ_{jk} and $\mu\langle\phi_j, g\rangle_{L^2(\partial\Omega)}$ to β_j. So J stays a convex quadratic in θ, its modified Gram matrix still positive semidefinite. A convex quadratic bounded below attains its minimum, so the modified normal equations are solvable, the same argument applies, and the limit minimizes $\Phi(u) = \|u - f\|^2 + \mu\|u - g\|^2_{L^2(\partial\Omega)}$ over V. This minimizer is unique, since $u \mapsto \|u - f\|^2$ is strictly convex and $u \mapsto \mu\|u - g\|^2_{L^2(\partial\Omega)}$ is convex, so Φ is strictly convex on V. Suppose the V_n become dense in $H^1(\Omega)$, that is $\operatorname{dist}_{H^1(\Omega)}(h, V_n) \to 0$ for every $h \in H^1(\Omega)$. The functions h_m of the proof of Theorem 3.2(b) satisfy $h_m \to f$ in $L^2(\Omega)$ and have trace g, so $\Phi(h_m) = \|h_m - f\|^2$. Fix m, and choose $v_n \in V_n$ with $v_n \to h_m$ in $H^1(\Omega)$. By the trace theorem $v_n \to h_m$ in $L^2(\partial\Omega)$ as well, so $\Phi(v_n) \to \Phi(h_m)$ and $\limsup_n \min_{V_n}\Phi \leq \|h_m - f\|^2$. Letting $m \to \infty$ gives $\min_{V_n}\Phi \to 0$. The limit u_{θ_∞} of the flow for the span V_n attains $\min_{V_n}\Phi$, so $\|u_{\theta_\infty} - f\|^2 \leq \min_{V_n}\Phi \to 0$ as $n \to \infty$. □

D Proofs for Consistency of Smooth Activations

Part (ii) of Theorem 6.1 assumes a stability estimate, which we prove first for the model problem.

Lemma D.1 (Stability of the model problem under boundary misfit)**.** *There is C_S, depending only on Ω, such that $\|w\| \leq C_S\left(\|-\Delta w + w\| + \|w\|_{L^2(\partial\Omega)}\right)$ for every $w \in C^2(\overline{\Omega})$.*

Proof. We argue by duality. Let $\psi \in L^2(\Omega)$ and let $v \in H^1_0(\Omega)$ be the weak solution of $-\Delta v + v = \psi$. Testing the weak form with v gives $\|\nabla v\|^2 + \|v\|^2 = \langle\psi, v\rangle \leq \|\psi\|\,\|v\|$, so $\|v\| \leq \|\psi\|$. Call the constant of Theorem 2.7 C; that fact then gives $v \in H^2(\Omega)$ with $\|v\|_{H^2(\Omega)} \leq 2C\|\psi\|$. For

$w \in C^2(\overline{\Omega})$, Green's second identity, valid for $v \in H^2(\Omega)$ by the density of $C^\infty(\overline{\Omega})$ in $H^2(\Omega)$ (Evans, 2010, Section 5.3, Theorem 3) and the continuity of both sides in the H^2 norm, reads

$$\int_\Omega (-\Delta w + w)\, v\, dx - \int_\Omega w\,(-\Delta v + v)\, dx = \int_{\partial\Omega} (w\partial_\nu v - v\partial_\nu w)\, dS = \int_{\partial\Omega} w\partial_\nu v\, dS,$$

since v vanishes on $\partial\Omega$. Hence $\langle w, \psi\rangle = \langle -\Delta w + w, v\rangle - \int_{\partial\Omega} w\partial_\nu v\, dS$, and by the trace theorem applied to $\nabla v \in H^1(\Omega)$, with a constant C' depending only on Ω,

$$\begin{aligned} |\langle w, \psi\rangle| &\le \|-\Delta w + w\|\, \|v\| + \|w\|_{L^2(\partial\Omega)}\, \|\nabla v\|_{L^2(\partial\Omega)} \\ &\le C' \left(\|-\Delta w + w\| + \|w\|_{L^2(\partial\Omega)}\right) \|v\|_{H^2(\Omega)}. \end{aligned}$$

Combining with $\|v\|_{H^2(\Omega)} \le 2C\,\|\psi\|$ and taking the supremum over $\|\psi\| = 1$ gives the claim with $C_S = 2CC'$. □

Proof of Theorem 6.1. Part (i). Let $U \in C^s(\mathbb{R}^d)$ restrict to $u^\star$, and let $\eta > 0$. By Theorem 2.5(i) with $k = 2$, applied to U, choose $u \in \mathcal{N}_\mathsf{S}$ with $\|u - u^\star\|_{W^{2,\infty}(\Omega)} \le \eta$. Using $\mathcal{L}u^\star = f$,

$$\|\mathcal{L}u - f\|_{L^\infty(\Omega)} = \|\mathcal{L}\,(u - u^\star)\|_{L^\infty(\Omega)} \le C_\mathcal{L}\eta,$$

so the interior residual is at most $|\Omega|\, C_\mathcal{L}^2\eta^2$. The difference $u - u^\star$ is continuous on $\overline{\Omega}$ with $|u - u^\star| \le \eta$, so $B(u) = \|u - u^\star\|^2_{L^2(\partial\Omega)} \le |\partial\Omega|\,\eta^2$. Summing gives $J(u) \le \left(|\Omega|^{1/2}\, C_\mathcal{L} + \mu^{1/2}\,|\partial\Omega|^{1/2}\right)^2 \eta^2$, and letting $\eta \to 0$ shows $\inf J = 0$. For the rate, fix a cube $Q \supset \overline{\Omega}$; then $U|_Q$ lies in $W^{s,\infty}(Q)$, say with norm M_0. Apply Theorem 2.5(ii) with $k = 2$ to U/M_0, and scale the output layer by M_0, which adds no weight. This gives networks with N nonzero weights and $\eta = O(N^{-(s-2-\varepsilon)/d})$, whence $J(u) = O(N^{-2(s-2-\varepsilon)/d})$, with constant $\left(|\Omega|^{1/2}\, C_\mathcal{L} + \mu^{1/2}\,|\partial\Omega|^{1/2}\right)^2 M_0^2 C^2$, where C is the constant of Theorem 2.5(ii).

Part (ii). Every $u \in \mathcal{N}_\mathsf{S}$ is smooth on $\overline{\Omega}$, and $u^\star \in C^2(\overline{\Omega})$. So the stability hypothesis of part (ii) applies to $w = u - u^\star$, and with $\mathcal{L}u^\star = f$ and $u^\star = g$ on $\partial\Omega$ it gives

$$\|u - u^\star\| \le C_S \left(\|\mathcal{L}u - f\| + \|u - g\|_{L^2(\partial\Omega)}\right) \le C_S \left(1 + \mu^{-1/2}\right) J(u)^{1/2},$$

since $\|\mathcal{L}u - f\|^2$ and $\mu\,\|u - g\|^2_{L^2(\partial\Omega)}$ are the two summands of $J(u)$ and each is at most $J(u)$. Along a minimizing sequence the right side tends to zero. For the networks of the rate in part (i), $J(u)^{1/2} = O\big(N^{-(s-2-\varepsilon)/d}\big)$, and the display turns this into the stated solution-error rate. □

Without the rate, the proof of Theorem 6.1 uses of the activation only derivative density and smoothness, and both extend beyond the saturating activations.

Corollary D.2 (Consistency for the sigmoid and Gaussian error linear units)**.** *Let $\sigma \in C^\infty(\mathbb{R})$ be non-polynomial with $\big\|\sigma^{(j)}\big\|_{L^\infty(\mathbb{R})} < \infty$ for every $j \ge 1$. Two examples are the sigmoid linear unit $t\varsigma(t)$, with ς the logistic sigmoid, and the Gaussian error linear unit $t\Phi(t)$, with Φ the standard normal distribution function. Write $\mathcal{N}_\sigma$ for the maps realizable with the single activation σ in the sense of Theorem 2.3. Then Theorem 6.1 holds with $\mathcal{N}_\sigma$ in place of $\mathcal{N}_\mathsf{S}$, except for its rate. That is, the infimum of J over $\mathcal{N}_\sigma$ is zero, and for $\mu > 0$ and under the stability estimate of part (ii) every minimizing sequence converges to $u^\star$ in $L^2(\Omega)$.*

Proof. First, the two examples satisfy the hypothesis. Write $\sigma(t) = t\psi(t)$ with $\psi \in \{\varsigma, \Phi\}$; by the Leibniz rule $\sigma^{(j)}(t) = j\psi^{(j-1)}(t) + t\psi^{(j)}(t)$ for $j \ge 1$. Every derivative $\psi^{(i)}$ is bounded. For ς, the identity $\varsigma' = \varsigma\,(1 - \varsigma)$ makes each derivative $\varsigma^{(i)}$ a polynomial in ς. Since ς takes values in $(0, 1)$, each such polynomial in ς is bounded, and so is $\varsigma^{(i)}$. For Φ, which takes values in $(0, 1)$, and $i \ge 1$, the derivative $\Phi^{(i)}$ is a polynomial times the standard normal density. For $i \ge 1$ the function $t\psi^{(i)}(t)$ is bounded as well. Indeed $\varsigma^{(i)}$ is $\varsigma\,(1 - \varsigma)$ times a polynomial in ς, and $\varsigma\,(1 - \varsigma) = e^{-t}\,(1 + e^{-t})^{-2} \le e^{-|t|}$; and a polynomial times the normal density decays faster than

$|t|^{-1}$. So $\sigma^{(j)}$ is bounded for every $j \geq 1$. Finally σ is smooth, and it is not a polynomial, since $\sigma(t) \to 0$ as $t \to -\infty$ while σ is not identically zero.

Second, derivative density. Put $\chi = \sigma'$. By hypothesis χ is smooth and bounded with every derivative, and it is not constant, since otherwise σ would be affine; a bounded non-constant function is not a polynomial, so $\chi \in \mathsf{S}$. Let $k \geq 0$, $G \in C^k(\mathbb{R}^d)$ and $\eta > 0$. Theorem 2.5(i), which is the theorem of Hornik (1991) for networks with one hidden layer, gives $v(x) = \sum_{l=1}^{m} a_l \chi\left(w_l \cdot x + \beta_l\right)$ with $\|v - G\|_{W^{k,\infty}(\Omega)} \leq \eta/2$. For $h > 0$ let

$$u_h(x) = \sum_{l=1}^{m} \frac{a_l}{h} \left(\sigma\left(w_l \cdot x + \beta_l + h\right) - \sigma\left(w_l \cdot x + \beta_l\right)\right),$$

a network with activation σ, one hidden layer and $2m$ units. For a multi-index α with $j = |\alpha| \leq k$, writing $t_l = w_l \cdot x + \beta_l$,

$$D^\alpha u_h(x) - D^\alpha v(x) = \sum_l a_l w_l^\alpha \left(\frac{\sigma^{(j)}\left(t_l + h\right) - \sigma^{(j)}\left(t_l\right)}{h} - \sigma^{(j+1)}\left(t_l\right)\right),$$

and by the mean value theorem the bracket equals $\sigma^{(j+1)}\left(\xi_l\right) - \sigma^{(j+1)}\left(t_l\right)$ for some ξ_l with $|\xi_l - t_l| \leq h$. As x ranges over $\overline{\Omega}$ each t_l ranges over a compact interval, on whose 1-neighborhood the continuous function $\sigma^{(j+1)}$ is uniformly continuous. So the bracket tends to zero uniformly on $\overline{\Omega}$ as $h \to 0$, and hence $\|u_h - G\|_{W^{k,\infty}(\Omega)} \leq \eta$ for h small.

Third, consistency. The proof of Theorem 6.1(i) uses of the activation only the density just proved, with $k = 2$, so the infimum is zero. Part (ii) uses of the activation only that every network is smooth on $\overline{\Omega}$, which holds for $\sigma \in C^\infty(\mathbb{R})$. □

E Further Related Work

Derivative-aware density and rates go back to Hornik and coauthors (Hornik, 1991; Hornik et al., 1994). Density without derivatives has the sigmoidal criterion of Cybenko (1989) and the non-polynomial criterion of Leshno et al. (1993). Rates are known for smooth activations (Mhaskar, 1996; Gühring & Raslan, 2021; De Ryck et al., 2021) and, in Sobolev norms, for the rectified linear unit (Gühring et al., 2020), and DeVore et al. (1989) give the nonlinear-width benchmark. For deep piecewise-polynomial activations the attainable derivative order grows with depth (Belomestny et al., 2023; Li et al., 2024), as does the network's degree (Li et al., 2020). What is special about degree one is that composing piecewise-linear maps stays piecewise linear, so depth cannot give a rectified-linear network a nonzero derivative of order two or more. A smooth non-polynomial activation has no finite degree, and the smooth saturating class has no wall, since Theorem 2.5(i) approximates every target together with its derivatives.

Theorem 3.2 also says why the convergence theory for physics-informed solvers in the literature needs its hypotheses: regularity control on the minimizing sequence (Shin et al., 2020), or an activation whose second derivative survives, such as the degree-three rectified power unit of Jiao et al. (2022). Activation choice also matters for weak-form objectives: in Deep Ritz computations for martensitic microstructures, Chen et al. (2023) use a smoothened rectified linear unit (SmReLU) to capture fine twinning and twin branching where the usual activations perform poorly. Rowan (2025) derive an autodifferentiation gradient error for a one-dimensional Deep Ritz objective with ReLU; that weak-form training effect differs from the strong-form residual collapse proved here. Such solutions lie outside the regularity Theorem 6.1 assumes. For trained weak forms, Rojas et al. (2024) pair a neural trial class with a fixed finite-dimensional test space and assume a local Fortin operator for that pair. We know of no construction of such an operator for a neural trial class.

E.1 Open Problems

Two problems seem to us both open and consequential. Theorem 5.3 settles the trajectory for fixed features, where the objective is quadratic in the trained weights. For descent on all the weights of a rectified-linear network we know of no guarantee that the risk tends to its infimum over the trained weights for a general source f, and no explicit rate. The natural tool is the Kurdyka–Łojasiewicz

theory for tame descent, which yields a rate only through a Łojasiewicz exponent. We know of no rule that carries such an exponent through the maxima and compositions a rectified-linear network is built from. Jentzen & Riekert (2022) apply that theory to deep rectified-linear regression under assumptions that include a piecewise-polynomial target function, and obtain a polynomial rate whose exponent they do not identify.

By Theorem 5.1 no penalty read at the almost-everywhere derivatives of a piecewise-linear network reads its curvature. Theorem C.1 records that a weak formulation, the conforming P_1 finite element method, is consistent on continuous piecewise-linear functions for the model problem with $g = 0$ on a polyhedral domain. Whether that consistency survives training with an inf-sup-stable pairing of trial and test classes is not known. Without such a pairing, zero-residual spurious minimizers can reappear (Berrone et al., 2022).