# Self-Consistent and Extremal Activation Functions: A Variational Theory of Activation Design

Anonymous

***Abstract*—Which activation function should a trained network have? We pose activation design as a variational problem on the affine Sobolev space $\sigma_{\text{base}} + H^2(\mathbb{R})$ — minimize empirical risk plus an $H^2$ penalty over $\sigma$, weights held fixed — and characterize the solutions structurally. Minimizers exist, and every critical point satisfies the *self-consistent representer equation* $\sigma^\star = \sigma_{\text{base}} - \frac{1}{\lambda} G * \bar{\mu}_{\sigma^\star}$, with $G$ the Green's function of $(1-\partial_u^2)^2$ and $\bar{\mu}_{\sigma^\star}$ the network's data-averaged perceptual measure: the optimal activation is a $C^2$ exponential spline whose knots are the pre-activation values the network itself produces under $\sigma^\star$ — not parameters, but a fixed point. Above an explicit threshold the problem is strongly convex on a ball containing every minimizer, the fixed point there is unique, and Picard iteration — gradient descent at step $1/\lambda$ — converges geometrically, at one backward pass per sample per iterate. An extremal strand explains piecewise-linearity: among all activations leaving every prediction of a trained network unchanged, the piecewise-linear interpolant minimizes the uniform first-order attack-surface bound, and under gradient fidelity the extremal has bang-bang curvature — a perfect spline in the sense of Favard and Karlin. Finally, parametric activation families (adaptive splines, Kolmogorov–Arnold edges) are Galerkin sections that $\Gamma$-converge to the full problem at the rate $O(h^{3/2})$. Every prediction is verified on trained MNIST networks over five seeds: the contraction exponent within $0.3\%$ of theory, knot self-consistency at grid resolution, the $h^{3/2}$ rate, and machine-precision behavioral equivalence of the piecewise-linear representative.**

***Index Terms*—activation functions, representer theorems, calculus of variations, splines, perfect splines, adversarial robustness, Gamma-convergence, Kolmogorov–Arnold networks**

## I. Introduction

Recent work [17] develops an analytic theory of how a trained feedforward network responds to perturbations of its activation function $\sigma$: the first-order response at input $x$ in output direction $v$ is integration against a finite signed atomic measure $\mu_{x,v}$ (the *perceptual measure*) supported at the network's pre-activation values, the worst-case response is its total variation, and the gradient flow of empirical risk on activation space is $\partial_t \sigma_t = -G * \bar{\mu}_{\sigma_t}$ with $G$ the Green's function of $(1-\partial_u^2)^2$. That theory takes $\sigma$ as given and asks how the network behaves near it.

This paper inverts the question: *which $\sigma$ should exist?* We treat the choice of activation for a trained network as a variational problem and characterize its solutions structurally — not by search, and not numerically, but by theorems that say what the optimizers look like. Two structural answers emerge, corresponding to two natural notions of optimality.

Under Hilbert-space (Tikhonov) regularization, the answer is a fixed point. Every critical point of the regularized risk satisfies

$$\sigma^\star \;=\; \sigma_{\text{base}} \;-\; \tfrac{1}{\lambda}\, G * \bar{\mu}_{\sigma^\star}, \qquad (1)$$

so $\sigma^\star - \sigma_{\text{base}}$ is a finite combination of kernels $G(\cdot - z)$ placed at the pre-activation values *that the network produces under $\sigma^\star$ itself*. The optimal activation is a spline; but unlike every learnable-activation method we know of, its knots are not free parameters — they are generated by the network's own forward pass, self-consistently. Because $G \in C^2(\mathbb{R})$ is an exponential kernel, the optimizer is a $C^2$ *exponential spline* in the sense of [18], exponentially localized to the network's operating range: the optimal activation differs from the base activation only where the network actually operates.

Under worst-case (extremal) optimality, the answer is a corner. Any two activations that agree at the finitely many operating points of a trained network induce *identical predictions on the data*; within this behavior class, the activation minimizing the uniform bound on the first-order attack surface of [17] is the piecewise-linear interpolant, and if one additionally demands gradient fidelity, the extremal has bang-bang second derivative — a perfect spline [7], [10]. To our knowledge this is the first extremal-problem explanation of why piecewise-linear (ReLU-family) nonlinearities and low-curvature smooth activations [16] occupy the special place they do: they are the extremals of a sensitivity functional.

*a) Contributions.:*

1) **A well-posed design problem (§III).** For a trained network with frozen weights, the regularized functional $J_\lambda(\sigma) = \mathcal{L}(\sigma) + \frac{\lambda}{2}\|\sigma - \sigma_{\text{base}}\|^2_{H^2(\mathbb{R})}$ attains its minimum on the affine space $\mathcal{A} = \sigma_{\text{base}} + H^2(\mathbb{R})$ for every $\lambda > 0$ (Theorem 5).
2) **The self-consistent representer theorem (§IV).** Critical points of $J_\lambda$ are exactly the solutions of (1) (Theorem 6); they are $C^2$ exponential splines with at most $nN$ knots ($n$ training samples, $N$ $\sigma$-neurons) located at their own pre-activations (Corollary 7), exponentially localized to the operating interval (Corollary 8). Above an explicit threshold $\lambda_0$, $J_\lambda$ is strongly convex on the relevant ball, the critical point there is unique and globally minimal, and Picard iteration — which is precisely gradient descent with step $1/\lambda$ — converges geometrically (Theorem 9, Proposition 10).
3) **Extremal activations of a trained network (§V).** Activations agreeing on the operating set of a trained network form a behavior class with identical data predictions (Lemma 12); the uniform attack-surface bound is monotone in $\|\sigma'\|_\infty$ (Proposition 13); its minimizer over the class is the piecewise-linear interpolant, with the extremal slope equal to the largest divided difference (Theorem 14); under

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.

value-and-derivative fidelity the extremal curvature is bang-bang (Theorem 16).

4) **The Galerkin hierarchy (§VI).** Parametric activation families are Galerkin sections of the variational problem; they $\Gamma$-converge to it, and above the threshold the restricted minimizers converge at the best-approximation rate, which is $O(h^{3/2})$ for spline families of mesh $h$ (Theorem 18, Corollary 19). Kolmogorov–Arnold networks [11] are the per-edge product instance.

*b) Relation to prior work.:* The nearest neighbor is Unser's representer theorem for deep networks [19] and its deep-spline implementation [2]: regularizing each neuron's activation by second-order total variation *during joint training* forces optimal activations to be piecewise-linear splines with adaptive knots (see also the Banach-space representer theorems of [14]). Our setting and conclusions differ in every load-bearing respect. The weights are fixed — design is post hoc, for a network that already exists; the geometry is Hilbert ($H^2$), not Banach ($\mathrm{TV}^2$), and accordingly produces $C^2$ exponential splines rather than piecewise-linear ones; and the knots are not free parameters of a representer formula but the solution of a fixed-point equation coupling the spline to the network's own forward pass. Piecewise-linearity, in our theory, is not the Tikhonov answer but the *extremal* one: it emerges from the worst-case sensitivity problem of §V, not from the regularized problem of §IV. The two regularities practitioners observe — spline-like learned activations and the robustness of low-curvature nonlinearities — thus arise here from two different variational principles, with the choice of regularizer acting as a design dial. Other lines are complementary: closed-form optimal activations in the asymptotic random-features regime [20], initialization-driven design [13], and search-based or parametric learnable activations [1], [15], [9], all of which our $\Gamma$-convergence result situates as sections of one function-space problem.

*c) Organization.:* §II imports the framework of [17] and fixes constants. §III–§VI develop the theory. §VII reports the experimental verification. §VIII closes with limitations and open problems. Proofs are collected in Appendix A; the Sobolev regularity computation behind the approximation rate is in Appendix B; experimental protocols are detailed in Appendix C.

## II. Preliminaries

Let $F_\sigma : \mathbb{R}^{d_0} \to \mathbb{R}^{d_L}$ be a feedforward network of depth $L$ with fixed weights and biases: pre-activations $z_\ell = W_\ell h_{\ell-1} + b_\ell$, post-activations $h_\ell = \sigma(z_\ell)$ componentwise for the $\sigma$-layers $\ell = 1, \dots, L-1$, and a linear read-out $F_\sigma(x) = z_L = W_L h_{L-1} + b_L$, with $h_0 = x$ — the standard classifier convention, and the one our experiments use. Write $N := \sum_{\ell=1}^{L-1} d_\ell$ for the number of $\sigma$-neurons; the perceptual measure below has its atoms at the $\sigma$-layers.

Following [17], the activation varies over the affine space

$$\mathcal{A} \;:=\; \sigma_{\mathrm{base}} + H^2(\mathbb{R}), \qquad \sigma_{\mathrm{base}} \in C^1_b(\mathbb{R}),$$

where $H^2(\mathbb{R}) = H^2(\mathbb{R})$ carries the inner product $\langle \eta_1, \eta_2 \rangle_{H^2(\mathbb{R})} = \langle (1-\partial_u^2)\eta_1, (1-\partial_u^2)\eta_2 \rangle_{L^2}$. The Riesz representative of point evaluation is the Green's function

$$G(u) = \tfrac{1}{4}(1+|u|)e^{-|u|}, \qquad (1-\partial_u^2)^2 G = \delta_0,$$

and $G \in C^2(\mathbb{R})$ with $G'(u) = -\frac{1}{4}ue^{-|u|}$ and $G''(u) = -\frac{1}{4}(1-|u|)e^{-|u|}$; the third derivative jumps at the origin.

For $x \in \mathbb{R}^{d_0}$, $v \in \mathbb{R}^{d_L}$, the *perceptual measure* is the finite signed atomic measure $\mu_{x,v} = \sum_{\ell,j} \lambda_\ell^j(x,v)\, \delta_{z_\ell^j(x)}$, where $\lambda_\ell = J_\ell^\top v$ are the adjoint states; it represents the Fréchet derivative of $\sigma \mapsto F_\sigma(x)$, and $\sup_{\|\eta\|_\infty \le 1} \langle v, DF_\sigma(x)[\eta] \rangle = \|\mu_{x,v}\|_{\mathrm{TV}}$ [17].

Throughout we make the following standing assumptions.

**Assumption 1.** (i) The training set $\{(x_i, y_i)\}_{i=1}^n$ is finite, with $\max_i \|x_i\|_\infty \le X < \infty$. (ii) The loss $\ell(\cdot, y)$ is nonnegative and $C^1$ with locally Lipschitz gradient. (iii) $\sigma_{\mathrm{base}} \in C^1_b(\mathbb{R})$. The weights $\{W_\ell, b_\ell\}$ are fixed.

The empirical risk and its data-averaged perceptual measure are

$$\mathcal{L}(\sigma) = \frac{1}{n}\sum_{i=1}^n \ell(F_\sigma(x_i), y_i), \quad \bar{\mu}_\sigma = \frac{1}{n}\sum_{i=1}^n \mu_{x_i, \nabla_F \ell_i}, \quad (2)$$

with $\nabla_F \ell_i := \nabla_F \ell(F_\sigma(x_i), y_i)$, so that $D\mathcal{L}(\sigma)[\eta] = \int \eta \, \mathrm{d}\bar{\mu}_\sigma$ and the $H^2(\mathbb{R})$-gradient of $\mathcal{L}$ at $\sigma = \sigma_{\mathrm{base}} + s$ is $G * \bar{\mu}_\sigma$ [17].

Three quantitative lemmas anchor everything that follows. The first makes the Morrey embedding explicit — with our normalization the constants are exactly $\frac{1}{2}$.

**Lemma 2** (Explicit embedding constants)**.** *For every $\eta \in H^2(\mathbb{R})$ and $z \in \mathbb{R}$,*

$$|\eta(z)| \le \tfrac{1}{2}\|\eta\|_{H^2(\mathbb{R})}, \qquad |\eta'(z)| \le \tfrac{1}{2}\|\eta\|_{H^2(\mathbb{R})}.$$

*Both constants are sharp.*

**Lemma 3** (Operating radius)**.** *For every $R > 0$ there is $Z_R < \infty$ such that for all $s \in H^2(\mathbb{R})$ with $\|s\|_{H^2(\mathbb{R})} \le R$ and all training inputs $x_i$, every pre-activation satisfies $\left| z_\ell^j(x_i;\, \sigma_{\mathrm{base}} + s) \right| \le Z_R$.*

**Lemma 4** (Gradient map on balls; [17])**.** *Define $\Phi_0(s) := G * \bar{\mu}_{\sigma_{\mathrm{base}}+s}$. For every $R > 0$ there are constants $M_R, L_R < \infty$ such that on the ball $B_R = \{\|s\|_{H^2(\mathbb{R})} \le R\}$,*

$$\|\Phi_0(s)\|_{H^2(\mathbb{R})} \le M_R := \tfrac{1}{2}\sup_{B_R} \|\bar{\mu}_{\sigma_{\mathrm{base}}+s}\|_{\mathrm{TV}},$$

*and $\Phi_0$ is Lipschitz on $B_R$ with constant $L_R$.*

Lemma 4 repackages the ingredients of the well-posedness proof for the activation gradient flow in [17]; Lemmas 2 and 3 are proved in Appendix A.

## III. The Variational Problem

Fix $\lambda > 0$ and define, for $\sigma = \sigma_{\mathrm{base}} + s \in \mathcal{A}$,

$$J_\lambda(\sigma) \;:=\; \mathcal{L}(\sigma) \;+\; \frac{\lambda}{2}\,\|s\|^2_{H^2(\mathbb{R})}. \quad (3)$$

This is Tikhonov regularization [6] in the geometry of [17]: the penalty is the squared distance from the base point in the $H^2(\mathbb{R})$ structure, so we ask for the best activation *near* the one the network was trained with, with $\lambda$ setting the exchange rate between risk and departure. Since $\ell \geq 0$, any minimizer satisfies the a priori bound

$$\|s^\star\|_{H^2(\mathbb{R})} \;\leq\; R_\lambda := \sqrt{2\mathcal{L}(\sigma_{\text{base}})/\lambda}, \tag{4}$$

obtained by comparing with $s = 0$. So the design problem is automatically local for large $\lambda$ and global for small $\lambda$.

**Theorem 5** (Existence)**.** *Under Assumption 1, $J_\lambda$ attains its minimum on $\mathcal{A}$ for every $\lambda > 0$, and every minimizer obeys* (4).

The proof (Appendix A) is the direct method with one twist worth naming: $H^2(\mathbb{R}) \hookrightarrow C^1_b(\mathbb{R})$ is *not* compact globally, but Lemma 3 confines all pre-activations to a fixed compact interval on bounded balls, and on compact intervals the embedding $H^2 \hookrightarrow C^1$ is compact by Rellich–Kondrachov [3]. Weak $H^2(\mathbb{R})$ convergence therefore moves the risk continuously, and the norm term is weakly lower semicontinuous.

The gradient of $J_\lambda$ in the tangent space $H^2(\mathbb{R})$ is

$$\nabla J_\lambda(s) \;=\; \lambda\, s \;+\; \Phi_0(s), \tag{5}$$

by (2) and Lemma 4. Two immediate consequences: $\nabla J_\lambda$ is Lipschitz on $B_R$ with constant $\lambda + L_R$, and — the engine of §IV — it is *strongly monotone* as soon as $\lambda$ exceeds $L_R$:

$$\begin{aligned}\langle \nabla J_\lambda(s_1) - \nabla J_\lambda(s_2),\, s_1 - s_2\rangle_{H^2(\mathbb{R})} \\ \geq\; (\lambda - L_R)\,\|s_1 - s_2\|^2_{H^2(\mathbb{R})}\end{aligned} \tag{6}$$

for $s_1, s_2 \in B_R$, i.e. $J_\lambda$ is $(\lambda - L_R)$-strongly convex there.

## IV. The Self-Consistent Representer Theorem

**Theorem 6** (Euler–Lagrange equation as a fixed point)**.** *$\sigma^\star = \sigma_{\text{base}} + s^\star \in \mathcal{A}$ is a critical point of $J_\lambda$ if and only if*

$$s^\star \;=\; -\frac{1}{\lambda}\, G * \bar{\mu}_{\sigma^\star}. \tag{7}$$

Equation (7) is a representer theorem with a loop in it. Because $\bar{\mu}_{\sigma^\star}$ is a finite atomic measure, the right-hand side is a finite superposition of kernels; but the atoms sit at the pre-activations generated by $\sigma^\star$ itself, so the equation determines the knots and the spline simultaneously. Unrolling the structure:

**Corollary 7** (Structure of critical points)**.** *Every critical point has the form*

$$\sigma^\star(u) \;=\; \sigma_{\text{base}}(u) \;-\; \frac{1}{\lambda}\sum_{a=1}^{K} c_a\, G(u - z_a), \tag{8}$$

*where $\{z_a\}_{a=1}^K$, $K \leq nN$, are the distinct pre-activation values of the network on the training set* under *$\sigma^\star$, and $c_a$ is the total adjoint weight of $\bar{\mu}_{\sigma^\star}$ at $z_a$. Consequently $\sigma^\star - \sigma_{\text{base}} \in C^2(\mathbb{R})$, and on each interval between consecutive knots it lies in $\operatorname{span}\{e^u, ue^u, e^{-u}, ue^{-u}\}$: the optimal correction is a $C^2$ exponential spline for the operator $(1-\partial_u^2)^2$ [18], with knots at the network's own operating points.*

**Corollary 8** (Localization)**.** *Let $[z_{\min}, z_{\max}]$ contain all knots and $d(u) := \operatorname{dist}(u, [z_{\min}, z_{\max}])$. Then*

$$|\sigma^\star(u) - \sigma_{\text{base}}(u)| \;\leq\; \frac{\|\bar{\mu}_{\sigma^\star}\|_{\text{TV}}}{4\lambda}\,\big(1 + d(u)\big)e^{-d(u)}.$$

*The optimal activation agrees with the base activation, up to exponentially small corrections, outside the network's operating range: optimality redesigns the nonlinearity only where the network actually computes.*

### A. Uniqueness and an algorithm from the contraction

Write $\Phi(s) := -\frac{1}{\lambda}\Phi_0(s)$ for the fixed-point map of (7).

**Theorem 9** (Unique self-consistent activation for large $\lambda$)**.** *Fix any radius $\bar{R} > 0$ and set*

$$\lambda_0 \;:=\; \max\{\, 2\mathcal{L}(\sigma_{\text{base}})/\bar{R}^2,\;\; M_{\bar{R}}/\bar{R},\;\; L_{\bar{R}}\,\}.$$

*For every $\lambda > \lambda_0$: (i) $\Phi$ maps $B_{\bar{R}}$ into itself and is a contraction with factor $q = L_{\bar{R}}/\lambda < 1$; (ii) $J_\lambda$ is $(\lambda - L_{\bar{R}})$-strongly convex on $B_{\bar{R}}$; (iii) the fixed point $s^\star \in B_{\bar{R}}$ is the unique critical point of $J_\lambda$ in $B_{\bar{R}}$ and is the global minimizer of $J_\lambda$ on $\mathcal{A}$; (iv) Picard iteration $s_{k+1} = \Phi(s_k)$ from any $s_0 \in B_{\bar{R}}$ satisfies $\|s_k - s^\star\|_{H^2(\mathbb{R})} \leq q^k\,\|s_0 - s^\star\|_{H^2(\mathbb{R})}$.*

**Proposition 10** (Picard is gradient descent)**.** *$\Phi(s) = s - \frac{1}{\lambda}\nabla J_\lambda(s)$: the fixed-point iteration of Theorem 9 is exactly gradient descent on $J_\lambda$ with step size $1/\lambda$.*

The iteration is implementable as stated. One step consists of a forward and backward pass per training sample (to assemble the atoms and adjoint weights of $\bar{\mu}$), followed by superposing $nN$ shifted copies of the closed-form kernel $G$ — the same primitive cost as one step of the activation gradient flow of [17]. The theory thus hands us a design algorithm with a certificate: above $\lambda_0$, geometric convergence at rate $L_{\bar{R}}/\lambda$ to the unique optimal activation. Below $\lambda_0$ the problem may cease to be convex; existence survives (Theorem 5), the representer structure survives at every critical point (Corollary 7), and the possible multiplicity of self-consistent activations becomes an experimental question (§VII).

### B. Symmetries and necessary conditions

The architecture carries exact symmetries that constrain every critical point. The shift family $\sigma \mapsto \sigma(\cdot - c)$, $b_\ell \mapsto b_\ell + c\mathbf{1}$ *for the $\sigma$-layers $\ell \leq L-1$* (the final bias is not shifted — shifting it would translate the output by $c\mathbf{1}$), leaves $F$ invariant for all $c \in \mathbb{R}$; differentiating at $c = 0$ yields a conservation identity in the style of Noether.

**Proposition 11** (Shift identity)**.** *For every $\sigma \in \mathcal{A}$,*

$$\int_{\mathbb{R}} \sigma'\,\mathrm{d}\bar{\mu}_\sigma \;=\; \sum_{\ell=1}^{L-1}\langle \nabla_{b_\ell}\mathcal{L}(\sigma),\, \mathbf{1}\rangle,$$

*where $\nabla_{b_\ell}\mathcal{L}(\sigma)$ is the bias gradient evaluated at that same $\sigma$. In particular, at any activation whose $\sigma$-layer biases are*

*loss-stationary — the trained activation itself if the biases were trained to convergence, or any $\sigma^\star$ after its biases are re-tuned to stationarity — the derivative of the activation is orthogonal to its own perceptual measure: $\int \sigma' \, \mathrm{d}\bar{\mu}_\sigma = 0$. (Bias-stationarity at $\sigma_{\mathrm{base}}$ does not transfer to a critical point $\sigma^\star \neq \sigma_{\mathrm{base}}$; the identity must be read at a single $\sigma$.)*

The dilation family $(x, b, \sigma) \mapsto (tx,\, tb,\, t\sigma(\cdot/t))$ similarly reproduces the Euler identity of [17] in averaged form; we do not restate it.

## V. Extremal Activations of a Trained Network

The Tikhonov problem of §IV asks for the best nearby activation. A sharper design question fixes the network's behavior exactly and minimizes its attack surface: among all activations that make the network compute *the same thing* on its data, which is the most robust? The answer turns out to be the classical extremal splines.

Fix $\sigma_0 \in C_b^1(\mathbb{R})$ (e.g. the trained activation) and a finite evaluation set $\mathcal{X}$ of inputs (training data, or training plus any inputs whose predictions must be preserved). Let $\mathcal{Z} = \{z_{(1)} < \cdots < z_{(K)}\}$ be the distinct pre-activation values of the network on $\mathcal{X}$ under $\sigma_0$, and $I := [z_{(1)}, z_{(K)}]$. Define the *behavior class* and *gradient class*

$$\mathcal{B}(\sigma_0) := \{\sigma \in \mathrm{Lip}_{\mathrm{loc}}(\mathbb{R}) : \sigma(z) = \sigma_0(z)\ \forall z \in \mathcal{Z}\},$$
$$\mathcal{B}^1(\sigma_0) := \{\sigma \in \mathcal{B}(\sigma_0) \cap C^1 : \sigma'(z) = \sigma_0'(z)\ \forall z \in \mathcal{Z}\}.$$

**Lemma 12** (Exact behavior invariance)**.** *Every $\sigma \in \mathcal{B}(\sigma_0)$ produces the same forward pass as $\sigma_0$ at every input in $\mathcal{X}$ — the same pre-activations, hidden states, and predictions — and hence the same operating set $\mathcal{Z}$. Every $\sigma \in \mathcal{B}^1(\sigma_0)$ additionally produces the same adjoint states and the same perceptual measures $\mu_{x,v}$ at every $x \in \mathcal{X}$.*

The proof is a one-line induction (Appendix A); it is the constructive face of the kernel characterization of $DF_\sigma$ in [17]: perturbations vanishing on $\operatorname{supp}\bar{\mu}$ are invisible to the network. What is *not* invariant on the class is the network's sensitivity at neighboring inputs — attacked inputs — whose pre-activations leave $\mathcal{Z}$. That is where the members of the class differ, and where extremality lives.

**Proposition 13** (The attack surface is monotone in the slope)**.** *Let $I' \supseteq I$ be an interval and $x$ any input whose pre-activations under $\sigma$ all lie in $I'$. Then*

$$\|\mu_{x,v}\|_{\mathrm{TV}} \;\le\; \|v\|_1 \sum_{\ell=1}^{L-1} \Big( \prod_{m=\ell+1}^{L} \|W_m\|_\infty \Big)\, \|\sigma'\|_{L^\infty(I')}^{L-1-\ell},$$

*where $\|W\|_\infty$ is the maximum absolute row sum. For networks with at least two $\sigma$-layers and nonzero weights, the bound is a strictly increasing polynomial in $\|\sigma'\|_{L^\infty(I')}$.*

Minimizing the uniform certificate over the behavior class is therefore the interpolation problem: minimize $\|\sigma'\|_{L^\infty(I)}$ subject to $\sigma = \sigma_0$ on $\mathcal{Z}$. Its solution is a corner stone of approximation theory.

**Theorem 14** (Piecewise-linear activations are extremal)**.** *Over $\sigma \in \mathcal{B}(\sigma_0)$,*

$$\min \|\sigma'\|_{L^\infty(I)} \;=\; s^\star := \max_{1 \le k < K} \frac{\big|\sigma_0(z_{(k+1)}) - \sigma_0(z_{(k)})\big|}{z_{(k+1)} - z_{(k)}},$$

*attained by the piecewise-linear interpolant of $(z_{(k)}, \sigma_0(z_{(k)}))$. Moreover $s^\star \le \|\sigma_0'\|_{L^\infty(I)}$, with equality only if $\sigma_0$ is affine on a maximizing gap.*

So: every trained network is *exactly* data-equivalent to a network with a piecewise-linear activation, and that representative minimizes the uniform first-order attack-surface bound of Proposition 13 over everything data-equivalent to it. Read as a design principle, this is an extremal-problem account of the ReLU family: piecewise-linearity is what worst-case sensitivity minimization produces once the network's behavior is pinned at its operating points. It is also the exact form of an industrial practice: *relufication* — swapping the smooth activations of a pretrained model for ReLU-type ones to exploit sparsity at inference, then fine-tuning to recover the lost accuracy [12] — replaces the activation wholesale; Lemma 12 says that interpolating it at the operating points instead makes the swap exact, with nothing to recover.

**Remark 15** (Extension off the operating interval)**.** Off $I$ the class $\mathcal{B}(\sigma_0)$ is unconstrained, while Proposition 13 is invoked on an interval $I' \supseteq I$ reached by attacked inputs. To make Theorem 14 govern the bound on $I'$, extend the interpolant beyond the extreme knots with any slope of modulus at most $s^\star$ — constant extension is the simplest choice — so that $\|\sigma'\|_{L^\infty(I')} = s^\star$ as well. Splicing back to $\sigma_{\mathrm{base}}$ instead (as in experiment E5) trades a little of the certificate outside $I$ for exact agreement with the base activation at infinity.

Demanding gradient fidelity as well — membership in $\mathcal{B}^1(\sigma_0)$, so that training gradients and perceptual measures on the data are also preserved — rules out corners, since a piecewise-linear function cannot match arbitrary derivative data. The extremal then migrates to the second derivative, and its structure is the classical perfect spline.

**Theorem 16** (Bang-bang curvature under gradient fidelity)**.** *The problem $\min \|\sigma''\|_{L^\infty(I)}$ over $\sigma \in \mathcal{B}^1(\sigma_0) \cap W^{2,\infty}(I)$ attains its minimum $\kappa^\star$, and among its minimizers there is a* perfect spline*: $|\sigma''| \equiv \kappa^\star$ almost everywhere on $I$ with finitely many sign changes [7], [10].*

The pair of theorems separates two design regimes that the literature usually conflates. Value fidelity alone drives the extremal to corners (ReLU-like); value-plus-gradient fidelity drives it to the smallest curvature budget compatible with the data, spent at full rate everywhere — a bang-bang profile whose smoothed corners are reminiscent of, and give an extremal rationale for, the empirically robust low-curvature activations of [16]. We emphasize the honest scope: Proposition 13 bounds the attack surface, and the extremal theorems minimize that bound; minimizing the exact total variation of the perceptual measure over the class is open (§VIII).

**Remark 17** (Two geometries, two splines)**.** The regularized problem (§IV) lives in the Hilbert space $H^2(\mathbb{R})$ and produces $C^2$ exponential splines; the extremal problem lives in $W^{1,\infty}/W^{2,\infty}$ and produces corners and bang-bang curvature. Unser's $\mathrm{TV}^2$ Banach regularization [19] sits between, producing piecewise-linear splines with free knots during training. The regularizer is a design instrument: it selects the spline species.

## VI. The Galerkin Hierarchy

Practical methods never optimize over all of $H^2(\mathbb{R})$; they optimize a parametric family. The variational problem organizes them. Let $V_1 \subset V_2 \subset \cdots \subset H^2(\mathbb{R})$ be closed subspaces with $\overline{\bigcup_m V_m} = H^2(\mathbb{R})$, and let $J_\lambda^m$ denote $J_\lambda$ restricted to $\sigma_{\mathrm{base}} + V_m$.

**Theorem 18** ($\Gamma$-convergence of Galerkin sections)**.** *Each $J_\lambda^m$ attains its minimum. The sequence $(J_\lambda^m)$ is equicoercive and $\Gamma$-converges to $J_\lambda$ in the weak topology of $H^2(\mathbb{R})$; consequently $\min J_\lambda^m \to \min J_\lambda$, and every weak cluster point of a sequence of restricted minimizers minimizes $J_\lambda$ [4].*

**Corollary 19** (Quantitative rate above the threshold)**.** *Let $\lambda > \lambda_0$ as in Theorem 9, $s^\star$ the unique minimizer and $s_m^\star$ the minimizer over $V_m$. Then*

$$\|s_m^\star - s^\star\|_{H^2(\mathbb{R})} \;\le\; \sqrt{\frac{\lambda + L_{\bar{R}}}{\lambda - L_{\bar{R}}}}\; \mathrm{dist}_{H^2(\mathbb{R})}(s^\star, V_m).$$

*For $V_m$ the compactly supported cubic B-spline spaces of mesh $h_m$ constructed in Appendix B, $\mathrm{dist}_{H^2(\mathbb{R})}(s^\star, V_m) = O(h_m^{3/2})$, because $s^\star$ lies in the Besov space $B^{7/2}_{2,\infty}(\mathbb{R})$ (and in $H^{7/2-\varepsilon}(\mathbb{R})$ for every $\varepsilon > 0$, but not $H^{7/2}$).*

The rate exponent $3/2$ is a fingerprint of the theory: it comes from the jump in $G'''$ at each self-consistent knot, i.e. from the Green's function of the $H^2$ geometry, not from generic smoothness assumptions. Instances of the hierarchy: parametric families such as PReLU [9] and adaptive piecewise-linear units [1] are low-dimensional sections (after smoothing into $H^2(\mathbb{R})$); learned spline activations [2] are exactly the spline sections above; Kolmogorov–Arnold networks [11] are the per-edge product of such sections, whose continuum limit is the tensorial version of (3) on a product of $H^2$ spaces — the natural setting for a sequel.

## VII. Experimental Results

The theory is quantitative and every constant in it is computable from one trained network, so its claims are directly falsifiable. We ran the program of Appendix C (frozen before execution) on five independently seeded MLPs (784–128–64–10, GELU base) trained on MNIST with early stopping; design-phase numerics are float64, and every $H^2$ norm is computed exactly through the reproducing property. Table I pairs each prediction with its measured outcome; bracketed intervals are 95% bootstrap confidence intervals over seeds.

Three results deserve comment. First, the *contraction law lands almost exactly*: regressing $\log q$ on $\log \lambda$ across the sweep gives slope $-0.998$ against the predicted $-1$, and the product $q\lambda$ is flat per seed (Fig. 1b) — Theorem 9(iv) observed as a measured law, not merely an upper bound. Second, *self-consistency*: in all 15 runs the knots of the converged $s^\star$ coincide with the pre-activation multiset the network produces under $\sigma^\star$, to the resolution $\delta$ of the measure grid (Fig. 1c) — the fixed-point signature of (7) that no free-knot spline method would reproduce. Third, the *rate*: the optimizer-free projection distance scales as $h^{1.54}$ against the predicted $h^{3/2}$ (Fig. 2), and the restricted-minimizer error tracks it within the predicted constant for $h \ge 0.2$; at the finest mesh the L-BFGS solution floor — not the theory — inflates the ratio to 1.56, and we report it as an optimizer artifact.

Two footnotes, both benign. The localization slope $-1.10$ is modestly *steeper* than $-1$: Corollary 8 is an upper envelope, and the signed atoms partially cancel in the far field, so the measured decay beats the worst-case bound — consistent, not anomalous. And uniqueness extends below the threshold: one limit cluster in every condition down to $\lambda_0/20$, so $\lambda_0$ — which carries a threefold safety factor on the sampled Lipschitz constant — is sufficient but empirically conservative.

Finally, the extremal strand behaves exactly as Lemma 12 and Theorem 14 say it must in the dense-knot regime: the piecewise-linear representative reproduces every evaluated prediction to machine precision, its slope factor equals 1.000, and its FGSM [8] robustness is statistically indistinguishable from the GELU original at every attack strength — exact behavioral equivalence with a strictly simpler activation, at no robustness cost.

## VIII. Discussion

*a) What the theory delivers.:* A design principle in three registers. Tikhonov: the best nearby activation is a self-consistent $C^2$ exponential spline, unique and computable by a certified fixed-point iteration for $\lambda$ above an explicit threshold. Extremal: the most robust activation with the network's exact data behavior is piecewise-linear, and bang-bang in curvature under gradient fidelity — an extremal-problem account of the ReLU family and of low-curvature design. Asymptotic: every parametric learnable-activation method is a Galerkin section of one function-space problem, with a convergence rate whose exponent is dictated by the $H^2$ Green's function.

*b) Limitations.:* The weights are frozen: this is post-hoc design, and the joint problem — weights and activation optimized together, the setting of [19] — couples the representer knots to weight dynamics in a way our fixed-point argument does not yet control. The extremal theorems minimize the uniform *bound* of Proposition 13; the extremal problem for the exact worst-case total variation is open, and we expect its solution to interpolate between the piecewise-linear and perfect-spline regimes. Below $\lambda_0$ the landscape of self-consistent activations is theoretically uncharted; our probes down to $\lambda_0/20$ found a unique limit throughout, so multiplicity and bifurcations in $\lambda$, if they occur, live at still smaller regularization — delimiting that regime is open. All

| Exp. | Tests | Prediction | Outcome (5 seeds) |
|---|---|---|---|
| E1 | Thm 9(iv) | geometric Picard convergence with $q < 1$; asymptotic ratio $\propto 1/\lambda$ | $q < 1$ and $J_\lambda$ descends in all 15 runs; geometric fits $R^2 = 0.99$ $[0.97, 1.00]$; slope of $\log q$ vs $\log\lambda = -0.998$ $[-1.031, -0.971]$ (pred. $-1$) |
| E2 | Cor. 7, 8 | knots of $s^\star$ = pre-activations under $\sigma^\star$; exponential decay off the operating interval | knot Hausdorff distance $0.0100 \le \delta = 0.02$ in $15/15$ runs; decay slope $-1.10$ $[-1.15, -1.05]$ (envelope bound $-1$; see text) |
| E3 | Thm 9(iii) | unique fixed point above $\lambda_0$; below exploratory | one limit cluster in all 20 (seed, $\lambda$) conditions, down to $\lambda_0/20$ |
| E4 | Cor. 19 | error $\frac{O(h^{3/2})}{\sqrt{(\lambda+L)/(\lambda-L)}}$ with constant $\le$ | projection-distance slope $1.54$ $[1.46, 1.63]$; minimizer-error slope $1.44$ $[1.35, 1.55]$; constant check holds for $h \ge 0.2$ (optimizer floor at $h = 0.1$; see text) |
| E5 | Lem. 12, Thm 14 | PL representative leaves evaluated predictions unchanged; bound rescales by exactly $s^\star / \|\sigma_0'\|_{L^\infty(I)}$ ($\approx 1$ for dense operating sets) | argmax agreement 1.0000, max logit deviation $3.6 \times 10^{-15}$; slope factor $= 1.000$ as predicted for dense $\mathcal{Z}$; FGSM accuracy change 0.0000 at every $\varepsilon$ |

TABLE I

PREDICTIONS VERSUS MEASURED OUTCOMES. EVERY CONSTANT IS COMPUTED FROM THE THEORY, NOT FIT; THE SHIFT IDENTITY OF PROPOSITION 11 ADDITIONALLY HOLDS WITH RELATIVE RESIDUAL $\le 7 \times 10^{-15}$.

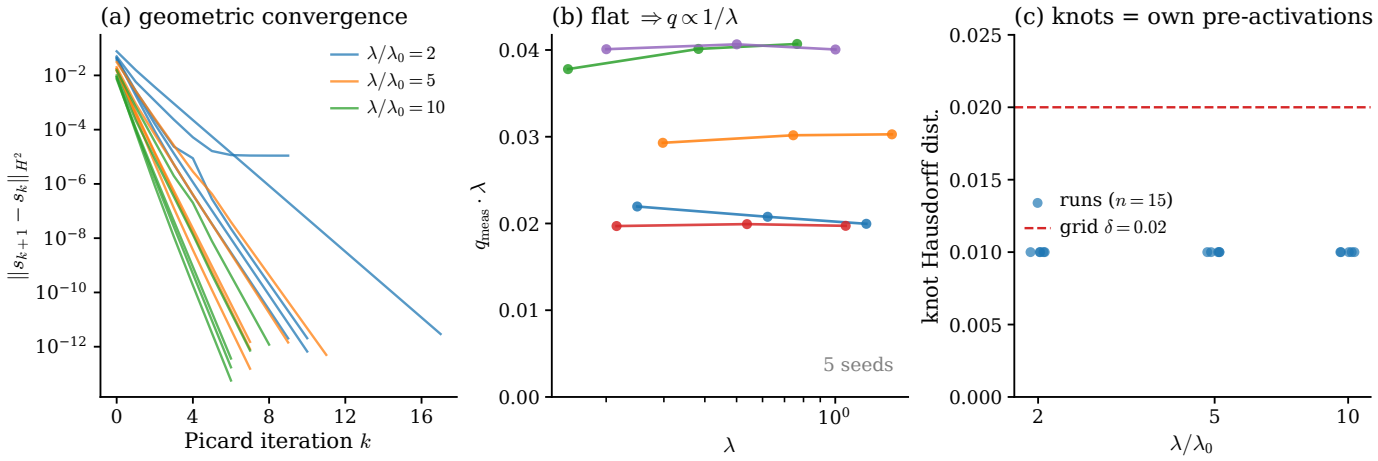


Fig. 1. **E1–E2 (Thm 9, Cor. 7).** (a) Picard steps $\|s_{k+1} - s_k\|_{H^2(\mathbb{R})}$ decay geometrically in every (seed, $\lambda$) run. (b) $q_{\text{meas}} \cdot \lambda$ is flat across the $\lambda$ sweep — the estimator-free signature of the contraction law $q = L/\lambda$. (c) the knots of the converged $s^\star$ coincide with the network's own pre-activations to the grid resolution $\delta$.

results are for a single shared $\sigma$; the per-neuron product space is the KAN limit and the natural sequel.

## APPENDIX A
## PROOFS

*Proof of Lemma 2.* Point evaluation: $\eta(z) = \langle G(\cdot - z), \eta\rangle_{H^2(\mathbb{R})}$ and $\|G(\cdot - z)\|^2_{H^2(\mathbb{R})} = \|G\|^2_{H^2(\mathbb{R})} = G(0) = \frac{1}{4}$, so $|\eta(z)| \le \frac{1}{2}\|\eta\|_{H^2(\mathbb{R})}$, with equality at $\eta = G(\cdot - z)$. Derivative evaluation: $(1 - \partial_u^2)^2\big(-G'(\cdot - z)\big) = -\delta_z'$, so $\langle -G'(\cdot - z), \eta\rangle_{H^2(\mathbb{R})} = -\delta_z'(\eta) = \eta'(z)$, and $\|G'\|^2_{H^2(\mathbb{R})} = \langle (1 - \partial^2)^2 G', G'\rangle_{L^2} = \delta_0'(G') = -G''(0) = \frac{1}{4}$; equality at $\eta = -G'(\cdot - z)$. □

*Proof of Lemma 3.* On $B_R$, $\|\sigma\|_\infty \le \|\sigma_{\text{base}}\|_\infty + \frac{R}{2} =: S_R$ by Lemma 2. Then $\|h_0\|_\infty = \|x_i\|_\infty \le X$ and $\|h_\ell\|_\infty \le S_R$ for $\ell \ge 1$, so $\left|z_\ell^j\right| \le \|W_\ell\|_\infty \max(X, S_R) + \|b_\ell\|_\infty$, and $Z_R$ is the maximum of these finitely many bounds over $\ell$. □

*Proof of Theorem 5.* Let $(s_k) \subset H^2(\mathbb{R})$ be a minimizing sequence for $j(s) := J_\lambda(\sigma_{\text{base}} + s)$. Since $\ell \ge 0$, $j(s) \ge \frac{\lambda}{2}\|s\|^2$, so $(s_k)$ is bounded; comparing with $s = 0$ we may assume $\|s_k\|^2 \le 2\mathcal{L}(\sigma_{\text{base}})/\lambda + 1 =: R^2$. Pass to a subsequence $s_k \rightharpoonup s_*$ weakly in $H^2(\mathbb{R})$; then $\|s_*\| \le \liminf \|s_k\|$.

We claim $\mathcal{L}(\sigma_{\text{base}} + s_k) \to \mathcal{L}(\sigma_{\text{base}} + s_*)$. By Lemma 3 all pre-activations of all $\sigma_{\text{base}} + s_k$ and of $\sigma_{\text{base}} + s_*$ lie in $[-Z_R, Z_R]$. Restriction $H^2(\mathbb{R}) \to H^2(-Z_R - 1, Z_R + 1)$ is bounded, and $H^2$ of a bounded interval embeds compactly in $C^1$ of its closure (Rellich–Kondrachov), so along a further subsequence $s_k \to s_*$ in $C^1([-Z_R - 1, Z_R + 1])$. Writing $\sigma_k = \sigma_{\text{base}} + s_k$, an induction over layers gives, for each training input,

$$\left\|h_\ell^{(k)} - h_\ell^{(*)}\right\| \le \text{Lip}(\sigma_*)\ \|W_\ell\|\ \left\|h_{\ell-1}^{(k)} - h_{\ell-1}^{(*)}\right\| + \varepsilon_k,$$

with $\varepsilon_k := \|\sigma_k - \sigma_*\|_{L^\infty[-Z_R, Z_R]} \to 0$; here every argument

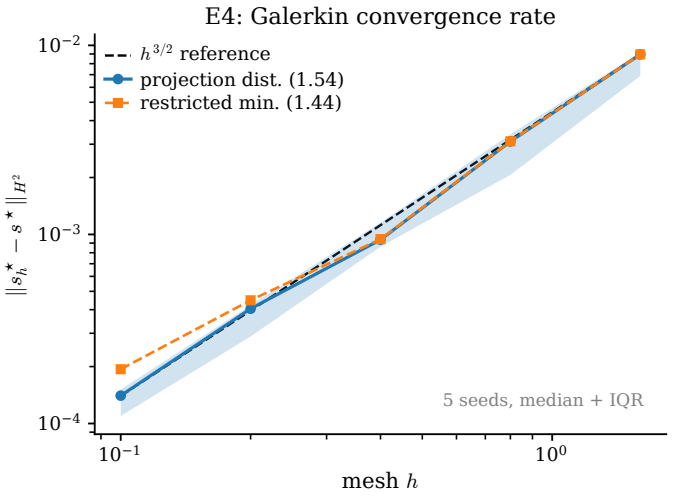


Fig. 2. **E4 (Cor. 19).** Galerkin convergence versus mesh $h$ on log–log axes (median + IQR over 5 seeds), with the $h^{3/2}$ reference. The optimizer-free projection distance gives slope 1.54 $[1.46, 1.63]$; the restricted-minimizer error tracks it for $h \geq 0.2$ and departs only at the finest mesh, where the L-BFGS solution floor — not the theory — flattens its overall slope to 1.44.

of every activation lies in $[-Z_R, Z_R]$ by Lemma 3, and $\mathrm{Lip}(\sigma_*)$ on that interval is finite since $\sigma_* \in C^1$. Hence $F_{\sigma_k}(x_i) \to F_{\sigma_*}(x_i)$ for each $i$, and continuity of $\ell$ gives the claim. Then

$$j(s_*) \leq \liminf_k \mathcal{L}(\sigma_{\text{base}} + s_k) + \tfrac{\lambda}{2} \liminf_k \|s_k\|^2 = \liminf_k j(s_k),$$

so $s_*$ is a minimizer. The a priori bound (4) follows for any minimizer by comparison with $s = 0$, using $\ell \geq 0$. □

*Proof of Theorem 6.* By (2) and the Riesz identification in [17], $DJ_\lambda(\sigma_{\text{base}} + s)[\eta] = \int \eta \, \mathrm{d}\bar{\mu}_{\sigma_{\text{base}}+s} + \lambda \langle s, \eta \rangle_{H^2(\mathbb{R})} = \langle \Phi_0(s) + \lambda s, \eta \rangle_{H^2(\mathbb{R})}$ for all $\eta \in H^2(\mathbb{R})$. Criticality is $\Phi_0(s^\star) + \lambda s^\star = 0$, which is (7). □

*Proof of Corollary 7.* $\bar{\mu}_{\sigma^\star} = \sum_a c_a \delta_{z_a}$ with $K \leq nN$ distinct atoms $z_a$ (coincident pre-activations merge), so (7) reads $s^\star = -\frac{1}{\lambda} \sum_a c_a G(\cdot - z_a)$, which is (8). Regularity: $G \in C^2(\mathbb{R})$, so $s^\star \in C^2$; away from the knots $(1 - \partial_u^2)^2 s^\star = 0$, whose solution space is $\mathrm{span}\{e^u, ue^u, e^{-u}, ue^{-u}\}$. □

*Proof of Corollary 8.* From (8), $|s^\star(u)| \leq \frac{1}{\lambda} \sum_a |c_a| \, G(u - z_a) \leq \frac{1}{\lambda} \|\bar{\mu}_{\sigma^\star}\|_{\text{TV}} \max_a G(u - z_a)$, and $\max_a G(u - z_a) \leq \frac{1}{4}(1 + d(u)) e^{-d(u)}$ since $G$ is even and decreasing in $|u|$. □

*Proof of Theorem 9.* (i) For $s \in B_{\bar{R}}$, $\|\Phi(s)\| \leq M_{\bar{R}}/\lambda \leq \bar{R}$ when $\lambda \geq M_{\bar{R}}/\bar{R}$; and $\|\Phi(s_1) - \Phi(s_2)\| \leq (L_{\bar{R}}/\lambda) \|s_1 - s_2\|$ by Lemma 4. Banach's fixed-point theorem gives a unique fixed point $s^\star \in B_{\bar{R}}$ and the geometric estimate (iv). (ii) is (6) with $R = \bar{R}$; on the convex set $B_{\bar{R}}$, strong monotonicity of the gradient is strong convexity. (iii) Critical points in $B_{\bar{R}}$ are exactly fixed points of $\Phi$ in $B_{\bar{R}}$ (Theorem 6), hence unique; by strong convexity $s^\star$ minimizes $J_\lambda$ over $B_{\bar{R}}$. Since $\lambda > 2\mathcal{L}(\sigma_{\text{base}})/\bar{R}^2$, the a priori bound (4) places every global minimizer (which exists by Theorem 5) inside $B_{\bar{R}}$, where the minimum is uniquely attained at $s^\star$. □

*Proof of Proposition 10.* $s - \frac{1}{\lambda} \nabla J_\lambda(s) = s - \frac{1}{\lambda}(\lambda s + \Phi_0(s)) = -\frac{1}{\lambda} \Phi_0(s) = \Phi(s)$. □

*Proof of Proposition 11.* For $c \in \mathbb{R}$ let $\sigma_c := \sigma(\cdot - c)$ and $b_\ell^c := b_\ell + c\mathbf{1}$ for $\ell \leq L-1$, with $b_L$ unchanged. Induction over the $\sigma$-layers: if $h_{\ell-1}$ is unchanged, then $z_\ell^c = W_\ell h_{\ell-1} + b_\ell + c\mathbf{1}$ and $\sigma_c(z_\ell^c) = \sigma(z_\ell^c - c) = \sigma(z_\ell) = h_\ell$. Hence $h_{L-1}$, and with it $z_L = W_L h_{L-1} + b_L$ and $F$, are unchanged (this is where the final bias must not be shifted), so $\mathcal{L}$ is constant along the family. Differentiating at $c = 0$: the $\sigma$-direction is $\partial_c \sigma_c|_{c=0} = -\sigma'$, contributing $D_\sigma \mathcal{L}[-\sigma'] = -\int \sigma' \, \mathrm{d}\bar{\mu}_\sigma$ (the representation theorem of [17] extended by continuity to $\sigma' \in C_b^0$, legitimate because $\bar{\mu}_\sigma$ is a finite atomic measure); the bias directions contribute $\sum_{\ell=1}^{L-1} \langle \nabla_{b_\ell} \mathcal{L}, \mathbf{1} \rangle$. The sum vanishes, giving the identity. The second statement sets the $\sigma$-layer bias gradients to zero. □

*Proof of Lemma 12.* Induction over layers at fixed input $x \in \mathcal{X}$: $h_0 = x$ is $\sigma$-independent; if $h_{\ell-1}$ agrees for $\sigma$ and $\sigma_0$, then $z_\ell$ agrees, its entries lie in $\mathcal{Z}$ by definition of $\mathcal{Z}$, and $\sigma = \sigma_0$ on $\mathcal{Z}$ gives $h_\ell$ agreement. For $\mathcal{B}^1$: the diagonals $D_\ell = \mathrm{diag}(\sigma'(z_\ell))$ then also agree, hence so do the Jacobians $J_\ell$, the adjoints $\lambda_\ell$, and the measures $\mu_{x,v}$. □

*Proof of Proposition 13.* Write $s := \|\sigma'\|_{L^\infty(I')}$. The linear read-out gives $\lambda_{L-1} = W_L^\top v$, so $\|\lambda_{L-1}\|_1 \leq \|W_L\|_\infty \|v\|_1$; for $\ell < L-1$ the adjoint recursion $\lambda_\ell = W_{\ell+1}^\top D_{\ell+1} \lambda_{\ell+1}$ gives $\|\lambda_\ell\|_1 \leq \|W_{\ell+1}\|_\infty \, s \, \|\lambda_{\ell+1}\|_1$ (the $1 \to 1$ operator norm of $W^\top$ is the maximum absolute row sum of $W$, and $\|D_{\ell+1}\|_{1\to 1} \leq s$ since all pre-activations lie in $I'$). Unrolling and summing over the $\sigma$-layers, with $\|\mu_{x,v}\|_{\text{TV}} \leq \sum_{\ell=1}^{L-1} \|\lambda_\ell\|_1$ from [17], yields the bound. Every summand except $\ell = L-1$ is a monomial of positive degree in $s$, so with at least two $\sigma$-layers and nonzero weight norms the sum is strictly increasing in $s$. □

*Proof of Theorem 14.* Lower bound: for any $\sigma \in \mathcal{B}(\sigma_0)$ locally Lipschitz and any $k$, the mean value inequality on $[z_{(k)}, z_{(k+1)}]$ gives $\|\sigma'\|_{L^\infty(I)} \geq |\sigma_0(z_{(k+1)}) - \sigma_0(z_{(k)})| / (z_{(k+1)} - z_{(k)})$; take the maximum over $k$. Attainment: the piecewise-linear interpolant lies in $\mathcal{B}(\sigma_0)$ and its slope on each gap is the corresponding divided difference, so its $L^\infty(I)$ slope is exactly $s^\star$. The comparison $s^\star \leq \|\sigma_0'\|_{L^\infty(I)}$ is the mean value theorem applied to $\sigma_0$, with equality forcing $|\sigma_0'| \equiv s^\star$ on a maximizing gap, i.e. affinity there. □

*Proof of Theorem 16 (sketch).* Existence: a minimizing sequence is bounded in $W^{2,\infty}(I)$; by weak-$*$ compactness and the compact embedding $W^{2,\infty}(I) \hookrightarrow C^1(\bar{I})$, a subsequence converges in $C^1$ to some $\sigma_\infty$ satisfying all value and derivative constraints, with $\|\sigma_\infty''\|_{L^\infty} \leq \liminf$ by weak-$*$ lower semicontinuity; hence the minimum $\kappa^\star$ is attained. Structure: the constrained problem is a minimal-$L^\infty$-norm interpolation of Hermite data, the extremal problem of Favard [7]; by the theory of generalized perfect splines [10] there is a minimizer whose second derivative has constant modulus $\kappa^\star$ and finitely many sign changes, the number bounded in terms of the

number of interpolation constraints. (When $\kappa^\star = 0$ the data are consistent with a single affine function and the statement is trivial.) □

*Proof of Theorem 18.* Existence of restricted minimizers: the proof of Theorem 5 runs verbatim inside the weakly closed subspace $V_m$. Equicoercivity: $J_\lambda^m \geq \frac{\lambda}{2}\|\cdot\|^2$ uniformly in $m$. $\Gamma$-liminf: if $s_m \rightharpoonup s$ then $\liminf J_\lambda(s_m) \geq J_\lambda(s)$ by the weak sequential lower semicontinuity established in the proof of Theorem 5 (the risk converges, the norm is wlsc). $\Gamma$-limsup: for $s \in H^2(\mathbb{R})$ take the recovery sequence $P_m s$ (orthogonal projection onto $V_m$); nestedness and density give $P_m s \to s$ strongly, hence also weakly, so the risk converges by the continuity established in the proof of Theorem 5, and the norm converges by strong convergence. The convergence of minima and of minimizers is the fundamental theorem of $\Gamma$-convergence [4]. □

*Proof of Corollary 19.* Let $P_m$ be the $H^2(\mathbb{R})$-orthogonal projection onto $V_m$ and abbreviate $L = L_{\bar{R}}$. Strong convexity (6) and $\nabla J_\lambda(s^\star) = 0$ give

$$\tfrac{\lambda - L}{2}\|s_m^\star - s^\star\|^2 \leq J_\lambda(s_m^\star) - J_\lambda(s^\star) \leq J_\lambda(P_m s^\star) - J_\lambda(s^\star),$$

the second inequality because $s_m^\star$ minimizes over $V_m \ni P_m s^\star$. The $(\lambda{+}L)$-Lipschitz gradient gives the quadratic upper bound $J_\lambda(P_m s^\star) - J_\lambda(s^\star) \leq \frac{\lambda+L}{2}\|P_m s^\star - s^\star\|^2$. Combine and take square roots; $\|P_m s^\star - s^\star\| = \mathrm{dist}_{H^2(\mathbb{R})}(s^\star, V_m)$. (For $m$ large all quantities lie in $B_{\bar{R}}$, where the constants are valid.) The spline rate is Appendix B. □

## Appendix B
## Sobolev regularity of the optimizer and the $O(h^{3/2})$ rate

The kernel has Fourier transform $\widehat{G}(\xi) = (1+\xi^2)^{-2}$, so

$$\|G\|_{H^s(\mathbb{R})}^2 = \int_{\mathbb{R}} (1+\xi^2)^s\,(1+\xi^2)^{-4}\,\mathrm{d}\xi < \infty \iff s < \tfrac{7}{2}.$$

By Corollary 7, $s^\star$ is a finite combination of shifted copies of $G$, hence $s^\star \in H^{7/2-\varepsilon}(\mathbb{R})$ for every $\varepsilon > 0$, and no better: the jump of $G'''$ at each knot obstructs $H^{7/2}$. The clean exponent comes from the Besov endpoint. Since $\left|\widehat{G}(\xi)\right|^2$ decays exactly like $|\xi|^{-8}$, the dyadic blocks $\int_{|\xi|\sim 2^j}(1+\xi^2)^{7/2}\left|\widehat{G}\right|^2$ are uniformly bounded, i.e. $G \in B^{7/2}_{2,\infty}(\mathbb{R})$, and so $s^\star \in B^{7/2}_{2,\infty}(\mathbb{R})$.

For the approximation space, take $V_h \subset H^2(\mathbb{R})$ to be the span of the cubic B-splines of mesh $h$ whose supports meet the window $[-A_h, A_h]$, with $A_h := Z + 2\log(1/h)$ and $Z$ the operating radius; these are compactly supported $C^2$ functions, so $V_h$ is a genuine finite-dimensional subspace of $H^2(\mathbb{R})$, and the family is nested and dense as $h \downarrow 0$ along dyadic meshes. Inside the window, quasi-interpolation of a $B^{7/2}_{2,\infty}$ function by cubic splines converges in $H^2$ at rate $h^{7/2-2} = h^{3/2}$ [5]; outside, Corollary 8 bounds $s^\star$ and its derivatives by a multiple of $(1{+}d)e^{-d}$, so the neglected tail contributes $O(e^{-(A_h - Z)}) = O(h^2)$ to the $H^2$ distance. Altogether

$$\mathrm{dist}_{H^2}(s^\star, V_h) \lesssim h^{3/2} + h^2 \lesssim h^{3/2},$$

which is the rate quoted in Corollary 19. The exponent is saturated: a single kernel $G(\cdot - z)$ with $z$ off the mesh already realizes it.

## Appendix C
## Experimental protocols

All experiments follow the experimental protocol of [17] (PyTorch; multi-seed with early stopping and plateau learning-rate decay; double precision for all kernel algebra). The protocols below were frozen before execution; §VII reports the outcomes. Trained models, per-run configurations, and per-iterate records are saved alongside the results for exact replication.

*a) E1 (contraction).:* Train MLPs (784–128–64–10, GELU base; 5 seeds) on MNIST; freeze weights. Estimate $L_{\bar{R}}$ by sampling $\|\Phi_0(s_1) - \Phi_0(s_2)\| / \|s_1 - s_2\|$ on random far and near pairs in $B_{\bar{R}}$; since sampled ratios only lower-bound the Lipschitz supremum, the threshold estimate $\widehat{\lambda}_0$ inflates $\widehat{L}$ by a safety factor of 3. Run Picard for $\lambda \in \{2, 5, 10\} \times \widehat{\lambda}_0$; record $\|s_{k+1} - s_k\|_{H^2(\mathbb{R})}$ per iterate (all $H^2$ norms exact via the reproducing property). The grid-binned map contracts to an $O(\delta)$ neighborhood of the exact fixed point rather than a point, so the iteration terminates at a step-size plateau of that scale; ratios are measured on the pre-plateau geometric phase. Predictions: geometric decay with asymptotic ratio $q < 1$, monotone descent of $J_\lambda$ (the Picard step is gradient descent at step $1/\lambda$, which descends whenever $\lambda \geq L$), and — the estimator-free signature of the contraction law — $q \cdot \lambda$ constant across the $\lambda$ sweep, equal to the local Lipschitz constant at $s^\star$. We also verify the shift identity of Proposition 11 at $\sigma_{\mathrm{base}}$ to numerical precision.

*b) E2 (self-consistency and localization).:* At convergence, extract the knot set of $s^\star$ and independently recompute the network's pre-activation multiset under $\sigma^\star$; report the two-sided Hausdorff distance. Prediction: at most the resolution $\delta$ of the grid on which the measure is represented (binning to width $\delta$ costs at most $(\delta/2)\,\|\bar{\mu}\|_{\mathrm{TV}}$ in $H^2$). Fit $\log\big(|s^\star(u)|\,/(1+d(u))\big)$ outside the operating interval against $d(u)$ (prediction: slope $-1$, Corollary 8).

*c) E3 (uniqueness and multiplicity).:* For $\lambda$ above and below $\widehat{\lambda}_0$, run damped gradient descent with the descent-safe step $1/(\lambda + \widehat{L})$ (Picard is the special case of step $1/\lambda$) from 10 random initializations in $B_{\bar{R}}$, each drawn from a per-initialization generator so runs replicate exactly; cluster the limits by pairwise $H^2$ distance. Prediction: one cluster above $\widehat{\lambda}_0$; below is exploratory (any multiplicity is a finding, not a failure).

*d) E4 (Galerkin rate).:* Solve the restricted problem on the cubic B-spline spaces of Appendix B with dyadic meshes $h \in \{1.6, 0.8, 0.4, 0.2, 0.1\}$, by full-batch L-BFGS on the coefficients, warm-started at the $H^2$ projection of $s^\star$ onto $V_h$ (legitimate: the restricted problem is strongly convex above the threshold, so its minimizer is unique and independent of initialization — the warm start only sharpens the optimizer to the precision the exact-distance comparison demands); the spline Gram is exact by Gauss–Legendre quadrature and the

cross terms with $s^\star$ are exact by the reproducing property. The projection distance itself is optimizer-free and is reported as the primary rate check. Predictions: log–log slope $3/2$, and the direct inequality $\|s_h^\star - s^\star\|_{H^2(\mathbb{R})} \le \sqrt{(\lambda + \widehat{L})/(\lambda - \widehat{L})} \cdot \operatorname{dist}_{H^2(\mathbb{R})}(s^\star, V_h)$, both sides computed exactly.

*e) E5 (piecewise-linear representative).:* With $\mathcal{X}$ = a fixed $10^4$-sample training subset plus the full test set, form the PL interpolant of the trained activation on $\mathcal{Z}$, splicing to $\sigma_{\text{base}}$ outside $[z_{(1)} - 1,\ z_{(K)} + 1]$ (cf. Remark 15). Verify prediction invariance on $\mathcal{X}$ to numerical precision — the interpolant is exact at the operating points by construction (Lemma 12); compute the sensitivity bound of Proposition 13 before and after (prediction: rescaling by exactly the factor $s^\star / \|\sigma_0'\|_{L^\infty(I)}$, which approaches 1 as the operating set densifies — a dense-knot regime we report honestly, alongside a sparse-$\mathcal{X}$ variant where the gap opens); measure FGSM [8] accuracy at the $\varepsilon$ grid of [17] before and after. The robustness delta is a measurement, not a prediction: the theorems certify the bound, and the experiment tests how much of the bound's behavior is realized empirically.

## REFERENCES


[1] F. Agostinelli, M. Hoffman, P. Sadowski, and P. Baldi. "Learning Activation Functions to Improve Deep Neural Networks." *ICLR Workshop*, 2015.

[2] P. Bohra, J. Campos, H. Gupta, S. Aziznejad, and M. Unser. "Learning Activation Functions in Deep (Spline) Neural Networks." *IEEE Open Journal of Signal Processing* 1 (2020): 295–309.

[3] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations.* Springer, 2011.

[4] G. Dal Maso. *An Introduction to* $\Gamma$*-Convergence.* Birkhäuser, 1993.

[5] R. A. DeVore and G. G. Lorentz. *Constructive Approximation.* Springer, 1993.

[6] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems.* Kluwer, 1996.

[7] J. Favard. "Sur l'interpolation." *Journal de Mathématiques Pures et Appliquées* 19 (1940): 281–306.

[8] I. J. Goodfellow, J. Shlens, and C. Szegedy. "Explaining and Harnessing Adversarial Examples." *ICLR*, 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." *ICCV*, 2015.

[10] S. Karlin. "Interpolation Properties of Generalized Perfect Splines and the Solutions of Certain Extremal Problems. I." *Transactions of the American Mathematical Society* 206 (1975): 25–66.

[11] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark. "KAN: Kolmogorov–Arnold Networks." *ICLR*, 2025.

[12] I. Mirzadeh, K. Alizadeh, S. Mehta, C. C. Del Mundo, O. Tuzel, G. Samei, M. Rastegari, and M. Farajtabar. "ReLU Strikes Back: Exploiting Activation Sparsity in Large Language Models." *ICLR*, 2024.

[13] M. Murray, V. Abrol, and J. Tanner. "Activation Function Design for Deep Networks: Linearity and Effective Initialisation." *Applied and Computational Harmonic Analysis* 59 (2022): 117–154.

[14] R. Parhi and R. D. Nowak. "Banach Space Representer Theorems for Neural Networks and Ridge Splines." *Journal of Machine Learning Research* 22.43 (2021): 1–40.

[15] P. Ramachandran, B. Zoph, and Q. V. Le. "Searching for Activation Functions." *Preprint arXiv:1710.05941*, 2017.

[16] V. Singla, S. Singla, S. Feizi, and D. Jacobs. "Low Curvature Activations Reduce Overfitting in Adversarial Training." *ICCV*, 2021, pp. 16423–16433.

[17] Anonymous. Under review, 2026. Author names withheld for double-blind review. "Perceptual Measures and Activation Flow: A Measure-Theoretic Theory of Sensitivity to Activation Functions in Neural Networks." *Preprint*, 2026.

[18] M. Unser and T. Blu. "Cardinal Exponential Splines: Part I—Theory and Filtering Algorithms." *IEEE Transactions on Signal Processing* 53.4 (2005): 1425–1438.

[19] M. Unser. "A Representer Theorem for Deep Neural Networks." *Journal of Machine Learning Research* 20 (2019): 1–30.

[20] J. Wang and J. Bento. "Optimal Activation Functions for the Random Features Regression Model." *Preprint arXiv:2206.01332*, 2022.