THE PRICE OF CONVEXITY IN SQUARED-LOSS REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

How much does requiring convex prediction losses restrict a model's representation? We study bounded scalar predictors on a fixed convex parameter domain, with squared losses that are ρ-weakly convex for every input and label. We prove that any finite prediction sign matrix realized with margin μ has sign-rank at most $q + 1$ whenever $\rho R^2 < 4\mu - 2\mu^2$, where q is the parameter dimension and R bounds the domain radius. The proof controls squared loss on convex hulls of parameter witnesses and then separates those hulls; it requires neither differentiability nor a nonvanishing gradient. An explicit family of bounded-weight ReLU networks with $O(d)$ nodes yields an exponential dimension separation: approximation within $\varepsilon < 1/2$ requires either $q + 1 \geq 2^{d/2}$ or $(1/2 + \varepsilon)^2 + \rho R^2/2 \geq 1$. A smooth two-parameter construction shows that the coefficient 4 in the small-margin curvature threshold is optimal. These are unconditional geometric representation bounds for direct squared-loss parameterizations, not lower bounds on optimization time or statistical performance.

1 INTRODUCTION

A model may represent useful functions compactly while its training objective is nonconvex. One response is to seek a different parameterization in which prediction loss is convex. But changing the parameterization can also change the number of coordinates needed to represent the same functions. This paper asks a specific geometric question:

> Can a low-dimensional parameterization retain a rich set of predictions while making every squared prediction loss convex, or nearly convex?

We study the *actual squared loss of a fixed prediction map*, not an arbitrary surrogate or a data-dependent lifted objective. This distinction matters: convex neural-network formulations can involve infinite-dimensional feature dictionaries, regularization, or training-set-dependent variables (Bach, 2017; Pilanci & Ergen, 2020). Their relationship to the prediction map must be checked before applying a representation bound.

Exact convexity is an informative but potentially brittle requirement. We therefore allow ρ-weak convexity: a loss becomes convex after adding $\rho \|\theta\|^2/2$. On a parameter domain of radius R, the scale-normalized quantity is ρR^2, not ρ alone. A rescaling of coordinates can make the latter arbitrarily small without changing predictions. We show that ρR^2 quantitatively controls the sign patterns a low-dimensional model can express at a prescribed output margin.

Results. Our main result converts a common weak-convexity bound for squared losses into a sign-rank bound (Theorem 2). A sign matrix realized at margin μ in q parameters has sign-rank at most $q+1$ whenever $\rho R^2 < 4\mu - 2\mu^2$. The argument works for nonsmooth prediction maps and arbitrary convex domains inside the radius bound. Combining it with an explicit ReLU construction and Forster's classical Hadamard sign-rank bound gives an exponential parameter-dimension separation (Theorem 5). Finally, a smooth two-parameter example exceeds the sign-rank limit while having weak-convexity certificate $4\mu + 8\mu^2$, establishing the optimal leading constant as $\mu \downarrow 0$ (Theorem 6). All three proofs are given in the main text.

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.

Position relative to prior work. Sign-rank measures representation by points and separating hyperplanes; its lower bounds and connections to learning are classical (Forster, 2002). More recently, Chornomaz et al. (2025) proved dimension lower bounds for reductions to stochastic convex optimization, including exponential separations from VC dimension. Thus, the observation that convex learning representations can require many coordinates is not itself our contribution. We instead give an explicit *margin–curvature threshold for direct bounded squared-loss models*, derive the corresponding neural representation bound, and determine the threshold's first-order small-margin constant. Our setting is narrower than general reductions to convex optimization, but it permits nonzero negative curvature.

Weak convexity is a standard framework for nonsmooth optimization (Davis & Drusvyatskiy, 2019); we use it to constrain representations, rather than to analyze an optimization algorithm. Convex composite-loss theory characterizes how loss functions and scalar links interact (Reid & Williamson, 2010). We impose no scalar-link form and obtain a finite sign-pattern consequence without proving that the predictions are affine or single-index. Quantitative neural approximation and depth-separation results motivate compact neural representations as a comparison point (Yarotsky, 2017; Telgarsky, 2016); our target construction and its size bound are explicit.

Our results preserve every sign at positive margin. They are not bounds for probabilistic embeddings or for representations allowed a nonzero fraction of classification errors. These distinctions are important in the work of Kamath et al. (2020) and, specifically for Hadamard matrices, in Bindua et al. (2026). We do not claim that all learning algorithms, all convex surrogates, or all implicit representations satisfy our hypotheses.

2 Setting and the role of the assumptions

Let $\mathcal{X}$ be an input set. A *direct parameterization* is a fixed map

$$h : \Theta \times \mathcal{X} \longrightarrow [-1, 1], \qquad \varnothing \neq \Theta \subseteq \mathbb{R}^q \text{ convex}, \tag{1}$$

written $h_\theta(x)$. The same domain and prediction map must serve every target in the family under consideration. The dimension q counts real parameter coordinates, not description bits, statistical capacity, or the size of an implicit dictionary. No continuity in x, differentiability in θ, or computational assumption is needed for the main theorem. The neural application uses continuous functions on $[-1, 1]^d$.

For labels $y \in [-1, 1]$, define

$$\ell_{x,y}(\theta) = (h_\theta(x) - y)^2, \qquad L_D(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell_{x_i,y_i}(\theta).$$

A function $f : \Theta \to \mathbb{R}$ is ρ-weakly convex if $f(\theta) + \rho \|\theta\|^2 /2$ is convex on Θ, where $\rho \geq 0$ (Davis & Drusvyatskiy, 2019). Convexity here is the usual inequality along all line segments in the domain; a Hessian need not exist. For a twice continuously differentiable function on an open convex set this is equivalent to $\nabla^2 f \succeq -\rho I$.

Proposition 1 (Endpoint losses suffice)**.** *The following are equivalent: all empirical losses L_D are ρ-weakly convex; all individual losses $\ell_{x,y}$ are ρ-weakly convex; and the two endpoint losses $\ell_x^{\pm} = (h_\theta(x) \mp 1)^2$ are ρ-weakly convex for every x.*

Proof. Singleton datasets give the first implication, and endpoint labels give the second. For the converse, use

$$(t - y)^2 = \frac{1 + y}{2}(t - 1)^2 + \frac{1 - y}{2}(t + 1)^2 - (1 - y^2). \tag{2}$$

The weights are nonnegative and sum to one. Adding $\rho \|\theta\|^2 /2$ preserves this convex combination, up to a constant. Averaging over a dataset preserves convexity as well. □

Parameter radius. Assume $\Theta \subseteq B(c, R)$, a closed Euclidean ball with $R > 0$. Replacing θ by $c + Ru$ gives weak-convexity parameter

$$\bar{\rho} = \rho R^2 \tag{3}$$

on a subset of the unit ball. Translation changes the quadratic correction only by an affine term. Thus $\bar{\rho}$ is unchanged by translating, rotating, or uniformly rescaling the coordinates with the corresponding change in radius. We do not assert invariance under general nonlinear reparameterization or arbitrary anisotropic preconditioning.

A constrained domain is meaningful rather than cosmetic. On all of $\mathbb{R}^q$, each bounded convex endpoint loss is constant: restricting to any line gives a convex function bounded above on $\mathbb{R}$, hence a constant. Since $\ell_x^- - \ell_x^+ = 4h_\theta(x)$, bounded predictions with globally convex endpoint losses would then be independent of θ. On constrained domains nonconstant examples are plentiful. If $\phi_j(x) \geq 0$ and $\sum_j \phi_j(x) = 1$, then $h_\theta(x) = \sum_j \theta_j \phi_j(x)$ with $\theta \in [-1, 1]^q$ is bounded and has convex squared losses. Nonlinear parameterizations are also possible (Appendix B).

Boundedness, uniformity, and margin. Bounded predictions ensure that a correctly signed output with margin μ has endpoint squared loss at most $(1-\mu)^2$; without an upper output bound this inference fails. We do not claim that clipping an arbitrary predictor preserves weak convexity. The requirement applies throughout the common parameter domain, not merely at training iterates or a fitted optimum. Finally, our margin is a margin in *prediction values*. It is not the normalized geometric margin of a linear classifier.

For a finite sign matrix $S \in \{-1, 1\}^{a \times b}$, its sign-rank is

$$\mathrm{sr}(S) = \min\{\mathrm{rank}(M) : M \in \mathbb{R}^{a \times b},\ S_{ij} M_{ij} > 0\ \forall i, j\}. \tag{4}$$

We say the model realizes S with margin $\mu > 0$ if there are inputs x_i and parameter witnesses $\theta_j \in \Theta$ such that $S_{ij} h_{\theta_j}(x_i) \geq \mu$ for all i, j. No distribution is selected in this definition.

3 A MARGIN–CURVATURE REPRESENTATION BOUND

Theorem 2 (Weak convexity forces low sign-rank)**.** *Let h be a direct parameterization on a convex domain $\Theta \subseteq B(c, R) \subseteq \mathbb{R}^q$, with bounded outputs and ρ-weakly convex endpoint squared losses for every input. If h realizes a finite sign matrix S with margin $0 < \mu \leq 1$, then*

$$\boxed{\bar{\rho} < 4\mu - 2\mu^2 \quad \Longrightarrow \quad \mathrm{sr}(S) \leq q + 1.} \tag{5}$$

Proof. Let x_i, θ_j be the witnesses, and fix row i. Partition the parameter witnesses into $P_i = \{\theta_j : S_{ij} = 1\}$ and $N_i = \{\theta_j : S_{ij} = -1\}$. For $z = \sum_j \lambda_j \theta_j \in \mathrm{conv}(P_i)$, where $\lambda_j \geq 0$ and $\sum_j \lambda_j = 1$, weak convexity gives

$$\begin{aligned} \ell_i^+(z) &\leq \sum_j \lambda_j \ell_i^+(\theta_j) + \frac{\rho}{2}\left(\sum_j \lambda_j \|\theta_j - c\|^2 - \|z - c\|^2\right) \\ &\leq (1-\mu)^2 + \frac{\rho R^2}{2} < 1. \end{aligned} \tag{6}$$

The final strict inequality is equivalent to the premise of (5). It implies $h_z(x_i) > 0$. The identical argument with ℓ_i^- shows $h_z(x_i) < 0$ for $z \in \mathrm{conv}(N_i)$. Therefore these two convex hulls are disjoint.

When both sets are nonempty, their convex hulls are compact because the witness sets are finite. They admit a strict affine separator, say $a_i^\top \theta + b_i$, positive on P_i and negative on N_i. A short finite-dimensional separation proof is included in Appendix A. If either sign is absent, use the corresponding nonzero constant separator. The matrix

$$M_{ij} = a_i^\top \theta_j + b_i$$

has signs S. It factors as the product of the matrix with rows $(a_i^\top, b_i)$ and the matrix with columns $(\theta_j^\top, 1)^\top$. Hence $\mathrm{rank}(M) \leq q + 1$. □

The key step is not linearity of the original predictor. It is that bounded squared loss keeps the positive and negative witness hulls disjoint. Only *after* establishing this property do we introduce affine separators. The resulting factorization reproduces the finite sign pattern, not necessarily the real-valued predictions.

Corollary 3 (Exact convexity)**.** *For a bounded direct parameterization with convex endpoint squared losses on any convex domain $\Theta \subseteq \mathbb{R}^q$, every finite strict prediction sign matrix has sign-rank at most $q+1$. No radius assumption is needed.*

Proof. Finitely many nonzero predictions have a positive minimum absolute value. Repeat the proof of Theorem 2 with $\rho = 0$; the quadratic term and the radius bound are unused. □

The strict threshold matters: the argument does not claim a sign-rank bound at equality. The contrapositive yields a necessary condition for a model exceeding sign-rank $q+1$,

$$\bar{\rho} \geq 4\mu - 2\mu^2. \tag{7}$$

Here ρ is any valid common weak-convexity certificate. A lower bound on it is not a lower bound on the runtime of an optimizer. Likewise, local Hessian measurements alone cannot establish the global premise of the theorem.

4 An explicit separation from compact ReLU representations

We now exhibit a continuous target family with small neural representations whose finite prediction matrix has large sign-rank. The Boolean vertices suffice for the lower bound; no claim about the interior behavior of a polynomial extension is needed.

Lemma 4 (Bounded ReLU parity extensions)**.** *For every $d \geq 1$, there are functions $g_A : [-1,1]^d \to [-1/2, 1/2]$, indexed by $A \subseteq [d]$, satisfying*

$$g_A(x) = \frac{1}{2} \prod_{j \in A} x_j \qquad (x \in \{-1,1\}^d). \tag{8}$$

Each has a ReLU computation graph with $O(d)$ nodes, depth $O(\log(d+1))$, weights from a fixed finite rational alphabet, and an explicit description of $O(d \log(d+1))$ bits.

Proof. Write $z_j = (1 - x_j)/2$. The gate

$$T(a,b) = \mathrm{ReLU}(a-b) + \mathrm{ReLU}(b-a) = |a-b|$$

maps $[0,1]^2$ into $[0,1]$ and computes XOR on Boolean inputs. Combine the z_j, $j \in A$, in a balanced binary tree and output $1/2 - T_A(z)$. For $A = \varnothing$, output $1/2$. A tree padded to a power of two with zero inputs has $O(d)$ gates and logarithmic depth; XOR with zero preserves a Boolean value and $T(a,0) = a$ for $a \in [0,1]$. Each gate uses a constant number of affine and ReLU nodes. Intermediate nonnegative values can be passed through ReLUs without change, giving the usual layered implementation with the same asymptotic bounds. All coefficients belong to a fixed rational alphabet of absolute value at most one. Node indices and wiring require $O(d \log(d+1))$ bits. The output is bounded on the entire cube and has the stated signs on its vertices. □

Theorem 5 (Neural representation separation)**.** *Suppose one direct q-parameter model with ρ-weakly convex squared losses on $\Theta \subseteq B(c,R)$ has, for every $A \subseteq [d]$, a witness θ_A satisfying*

$$\sup_{x \in [-1,1]^d} |h_{\theta_A}(x) - g_A(x)| \leq \varepsilon < 1/2.$$

Then

$$\boxed{q + 1 < 2^{d/2} \quad \Longrightarrow \quad (\tfrac{1}{2} + \varepsilon)^2 + \frac{\bar{\rho}}{2} \geq 1.} \tag{9}$$

In particular, exact convexity requires $q \geq 2^{d/2} - 1$. The same conclusions hold if approximation is required only on the Boolean vertices.

Proof. Set $N = 2^d$. The matrix of signs of (8), with rows indexed by Boolean vertices and columns by subsets, is a Walsh–Hadamard matrix H_d. Its columns are the parity characters, so

$$H_d H_d^\top = NI, \qquad \|H_d\|_{\mathrm{op}} = \sqrt{N}.$$

Forster's spectral lower bound gives $\mathrm{sr}(H_d) \geq N / \|H_d\|_{\mathrm{op}} = \sqrt{N}$ (Forster, 2002). Approximation supplies witnesses with margin $\mu = 1/2 - \varepsilon$. If $q + 1 < \sqrt{N}$, Theorem 2 forces $(1-\mu)^2 + \bar{\rho}/2 \geq 1$, which is (9). The exact-convex case also follows directly from Corollary 3. □

For example, approximation error $\varepsilon = 1/4$ and $q + 1 < 2^{d/2}$ require $\bar{\rho} \geq 7/8$. More generally, every polynomial-dimensional sequence of such parameterizations faces the bound for all sufficiently large d. This is a separation between a *shared direct parameterization* and a family of individually compact neural representations, not between the bit lengths of individual predictors. The target index A may be encoded in a neural architecture; the alternative must use one fixed domain and map to represent all targets.

The approximation tolerance is substantive. At error $1/2$, the constant-zero predictor approximates every g_A and has convex losses. The family-level quantifier is also essential: for each small admissible model some target is not approximated. A single fixed target could instead be built into a constant parameterization. Finally, the construction and Forster bound are elementary or established ingredients; the use of the margin–curvature theorem supplies the nearly convex representation conclusion.

5 The optimal leading constant at small margins

The next construction addresses the universal threshold in Theorem 2, not the tightness of the exponential dimension exponent for the neural family.

Theorem 6 (Small-margin sharpness)**.** *For every $0 < \mu \leq 1/2$, there is a smooth direct model on the unit disk $\Theta = B_2(0, 1)$, with inputs $x \in [-1, 1]^4$ and outputs in $[-1, 1]$, that realizes a sign matrix of sign-rank $4 > q + 1 = 3$ at margin μ. All of its squared losses for $y \in [-1, 1]$ are $(4\mu + 8\mu^2)$-weakly convex. Consequently, the leading coefficient 4 in (7) cannot be increased.*

Proof. For $\theta = (u, v)$ define

$$\begin{aligned} h_{(u,v)}(x) &= \mu\{a_x + b_x u + c_x v + d_x(u^2 - v^2)\}, \qquad (10)\\ a_x &= (x_1 + x_2 + x_3 + x_4)/4, \qquad b_x = (x_1 - x_2)/2,\\ c_x &= (x_3 - x_4)/2, \qquad d_x = (x_1 + x_2 - x_3 - x_4)/4. \end{aligned}$$

At the parameter witnesses $e_1, -e_1, e_2, -e_2$, the four predictions are $\mu x_1, \mu x_2, \mu x_3, \mu x_4$. Therefore, using all 16 Boolean inputs gives a 16×4 sign matrix S_* containing every sign vector on four columns.

This matrix has sign-rank four. Indeed, a real matrix M with these signs and rank less than four would have a nonzero kernel vector $z \in \mathbb{R}^4$. Choose a row whose signs agree with z on all nonzero coordinates. Each nonzero product $M_{ij} z_j$ is then positive, so that row of Mz cannot vanish. The reverse inequality is immediate from the four columns.

We next bound the predictions uniformly. For a Boolean input with zero or four positive coordinates, the polynomial inside braces in (10) is constant ± 1. With one or three positive coordinates, its coefficients satisfy $|a_x| = |d_x| = 1/2$ and $|b_x| + |c_x| = 1$, so its absolute value on the disk is at most two. With two positive coordinates, either $a_x = d_x = 0$ and $|b_x| = |c_x| = 1$, giving bound $\sqrt{2}$, or $a_x = b_x = c_x = 0$ and $|d_x| = 1$, giving bound one. Since the prediction is affine in x, its absolute maximum over the input cube is attained at a vertex. Hence $|h_{(u,v)}(x)| \leq 2\mu \leq 1$ everywhere.

Finally, $|d_x| \leq 1$ gives

$$\nabla^2_\theta h = 2\mu d_x \operatorname{diag}(1, -1), \qquad \left\|\nabla^2_\theta h\right\|_{\mathrm{op}} \leq 2\mu.$$

For every input, label, and point in the disk,

$$\begin{aligned} \nabla^2_\theta (h - y)^2 &= 2\nabla h \nabla h^\top + 2(h - y)\nabla^2 h\\ &\succeq -4\mu(1 + 2\mu) I. \qquad (11) \end{aligned}$$

The polynomial is smooth on a neighborhood of the disk; the Hessian inequality along its line segments proves the claimed weak convexity, including at boundary points. Since $R = 1$ and

$$\frac{4\mu + 8\mu^2}{4\mu - 2\mu^2} \longrightarrow 1 \qquad (\mu \downarrow 0),$$

any strictly larger universal leading coefficient would contradict this construction for sufficiently small μ. □

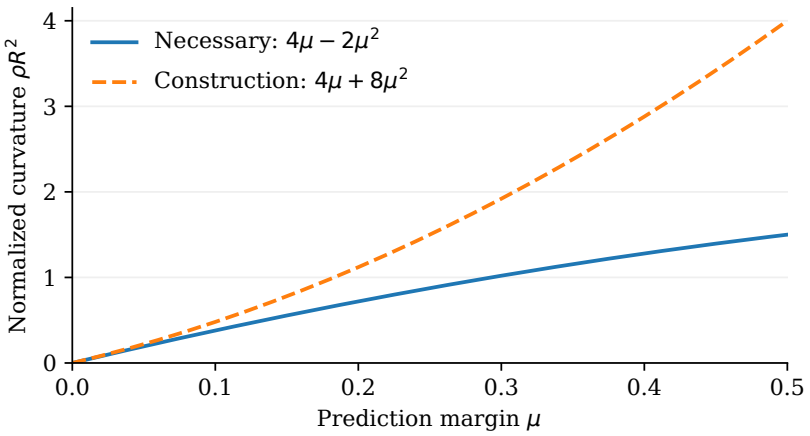

Figure 1: Analytic lower and upper bounds for violating $\mathrm{sr}(S) \leq q + 1$. The necessary bound is $4\mu - 2\mu^2$; the two-parameter construction admits certificate $4\mu + 8\mu^2$. Their leading small-margin coefficient agrees. The upper curve is a sufficient certificate, not an exact minimal-curvature calculation. No optimization experiment is plotted.

To state the sharpness precisely, let $r_*(\mu)$ be the infimum of ρR^2 over all models and finite sign matrices satisfying the hypotheses of Theorem 2 except its numerical threshold, with margin at least μ and sign-rank exceeding $q + 1$. For $0 < \mu \leq 1/2$, Theorems 2 and 6 imply

$$4\mu - 2\mu^2 \leq r_*(\mu) \leq 4\mu + 8\mu^2, \qquad \lim_{\mu \downarrow 0} \frac{r_*(\mu)}{\mu} = 4. \tag{12}$$

This limit does not identify $r_*(\mu)$ at finite margin. It also does not construct a polynomial-dimensional, low-curvature parameterization of the entire parity family.

6 Scope and Conclusion

The results isolate a representation consequence of requiring every squared prediction loss to have small negative curvature. They apply even without smoothness or an optimization algorithm, but only on a shared convex domain for a direct prediction map. Convexity of a regularized training objective, a surrogate loss, or a data-dependent lift alone does not establish this premise. In particular, our results do not contradict the settings of Bach (2017); Pilanci & Ergen (2020).

The charged resource is explicit real-coordinate dimension. Exponential dimension does not by itself imply exponential encoded size for every sparse or implicit predictor, poor statistical performance, or slow evaluation. A lower bound on negative curvature likewise does not imply that the corresponding nonconvex objective is difficult to optimize. The theorem is deterministic and pointwise: transferring the conclusion to distribution-dependent approximation or randomized representations would require a different argument. The distinction is particularly important for the Hadamard example, whose exact and approximate sign-ranks can differ substantially (Kamath et al., 2020; Bindua et al., 2026).

The sharpness result identifies the first-order curvature threshold, not a complete dimension–curvature frontier. Determining the finite-margin frontier or finding dimension-matched upper constructions are concrete remaining geometric questions. The radius and bounded-output assumptions cannot simply be dropped: the former controls interpolation across parameter witnesses, while the latter turns output margin into a squared-loss bound. Appendix B records illustrative boundary cases.

In summary, sufficiently small scale-normalized negative curvature forces the finite signs of a bounded squared-loss model to admit a low-dimensional affine representation. Compact ReLU target families can violate that requirement exponentially in input dimension. Allowing a controlled departure from convexity changes the conclusion, and the universal small-margin coefficient governing that departure is exactly four.

AI USE STATEMENT

Generative AI was used extensively to develop and refine the mathematical formulation, formulate claims and proof arguments, assist in proof checking and literature discovery, and draft and edit the manuscript. It was also used to implement consistency checks, generate the analytic figure, and prepare the LaTeX artifact. The reported checks consist of the written derivations, source comparisons, exact finite calculations, symbolic identities, and build tests described in the supplement. They are not proof-assistant verification or an independent expert review. No empirical training results or human-subject data are generated or reported.

REPRODUCIBILITY AND ETHICS STATEMENTS

The core proofs appear in the main text. The supplement contains exact checks for the finite constructions and symbolic identities, figure-generation code with rational source data, and build instructions. The figure displays proved formulas rather than fitted measurements. This theoretical study involves no personal data, human participants, or deployed system. Its geometric conclusions should not be used to infer the reliability or computational difficulty of practical learning systems outside the stated assumptions.

REFERENCES

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017. URL https://jmlr.org/papers/v18/14-546.html.

Riju Bindua, Hamed Hatami, Hasti Karimi, and Robert Robere. Lower bounds for approximate sign rank. arXiv preprint arXiv:2605.01038, 2026. URL https://arxiv.org/abs/2605.01038.

Bogdan Chornomaz, Shay Moran, and Tom Waknine. On reductions and representations of learning problems in euclidean spaces. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 2043–2054. ACM, 2025. doi: 10.1145/3717823.3718213. URL https://arxiv.org/abs/2411.10784.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019. doi: 10.1137/18M1178244. URL https://arxiv.org/abs/1803.06523.

Jürgen Forster. A linear lower bound on the unbounded error probabilistic communication complexity. *Journal of Computer and System Sciences*, 65(4):612–625, 2002. doi: 10.1016/S0022-0000(02)00019-3.

Pritish Kamath, Omar Montasser, and Nathan Srebro. Approximate is good enough: Probabilistic variants of dimensional and margin complexity. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2236–2262, 2020. URL https://proceedings.mlr.press/v125/kamath20b.html.

Mert Pilanci and Tolga Ergen. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7695–7705, 2020. URL https://proceedings.mlr.press/v119/pilanci20a.html.

Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010. URL https://jmlr.org/papers/v11/reid10a.html.

Matus Telgarsky. Benefits of depth in neural networks. In *Proceedings of the 29th Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1517–1539, 2016. URL https://proceedings.mlr.press/v49/telgarsky16.html.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017. doi: 10.1016/j.neunet.2017.07.002.

A Auxiliary details for the geometric proof

For completeness, finite convex hulls $C_+, C_- \subseteq \mathbb{R}^q$ that are nonempty and disjoint admit strict affine separation. Both are compact, so there are closest points $p \in C_+$ and $n \in C_-$ with $\delta = \|p - n\| > 0$. Set $w = p - n$. Minimizing squared distance along line segments gives

$$\langle w, p' - p \rangle \geq 0 \quad (p' \in C_+), \qquad \langle w, n' - n \rangle \leq 0 \quad (n' \in C_-).$$

Consequently the affine map $z \mapsto \langle w, z - (p+n)/2 \rangle$ is at least $\delta^2/2$ on C_+ and at most $-\delta^2/2$ on C_-. This proves the separation step used in Theorem 2. Closedness of the full parameter domain is unnecessary: the two finite convex hulls are compact and remain inside it by convexity.

The weak-convexity correction in (6) has the variance identity

$$\sum_j \lambda_j \|\theta_j - c\|^2 - \|z - c\|^2 = \sum_j \lambda_j \|\theta_j - z\|^2, \qquad z = \sum_j \lambda_j \theta_j.$$

It is nonnegative and at most R^2 for points in $B(c, R)$. In particular, no factor depending on the number of witnesses appears. The case of an empty positive or negative witness set is handled by a constant separator and needs no compactness argument. If the domain has a smaller affine hull, the argument may be applied in that affine hull with its inherited Euclidean metric; no such reduction is required for the stated bound.

B Boundary examples and interpretation

Nonlinear predictions with convex squared loss. Convex endpoint losses do not require affine predictions. Let $f(t) = t + t^2/10$ on $[-1/2, 1/2]$. Then $|f| \leq 0.525$, $f' \in [0.9, 1.1]$, and $f'' = 1/5$. For every $y \in [-1, 1]$,

$$\frac{1}{2}\frac{d^2}{dt^2}(f(t) - y)^2 = (f')^2 + (f - y)f'' \geq \frac{81}{100} - \frac{1}{5}\frac{59}{40} = \frac{103}{200} > 0.$$

Here the lower bound uses $f(t) \geq -19/40$. The model is nonlinear but satisfies exact convexity; it does not challenge Corollary 3, which concerns signs rather than affine prediction values.

Why one endpoint is not enough. A nonlinear predictor can make one endpoint loss constant or convex while the opposite endpoint has negative curvature. For example, on $[-1/2, 1/2]$ let $h_t = t^2$. Then $(h_t + 1)^2 = (t^2 + 1)^2$ is convex, whereas the second derivative of $(h_t - 1)^2 = (t^2 - 1)^2$ equals $12t^2 - 4$, which is negative near zero. The two endpoint conditions in Proposition 1 therefore cannot be replaced by one.

Why radius cannot be ignored. Apply the change of variables $\theta = au$ to any nonconstant example on a unit ball. The new domain radius is a, while a valid weak-convexity certificate is divided by a^2. Taking a large makes this certificate small without changing the set of predictions, their margins, or the sign matrix. The product ρR^2 remains unchanged. More general coordinate changes may alter the geometry; they are not covered by this normalization argument.

Why an infimum need not be attained. The separation result assumes actual approximation witnesses. If instead one knows $\inf_\theta \|h_\theta - g_A\|_\infty \leq \varepsilon$, use witnesses with error at most $\varepsilon + \delta < 1/2$, prove (9) with that error, and let $\delta \downarrow 0$. The finite number of targets for each d permits one common δ. Thus the same necessary inequality holds in the infimum formulation without assuming a minimizer exists.

C Construction checks and reproducibility

The symbolic and finite tests in the supplement verify identities and examples used in the proofs. They do not constitute a proof of a universally quantified theorem or an exhaustive novelty check.

ReLU witnesses. The script checks the XOR gate on all four Boolean inputs, evaluates every vertex–subset pair for dimensions $d = 1, \ldots, 8$, and verifies $H_d H_d^\top = 2^d I$ with integer arithmetic. This gives $\sum_{d=1}^{8} 4^d = 87{,}380$ exact prediction checks. It also evaluates a grid of rational inputs to check boundedness of the continuous gate composition. The proof of boundedness for all inputs is the gate-wise induction in Lemma 4, not the finite grid.

Sharpness witnesses. The script verifies the four interpolation identities of (10) symbolically, classifies all sixteen input vertices into the coefficient cases in the proof, and checks the polynomial squared-loss Hessian identity. It verifies the kernel-sign contradiction on a finite set of representative vectors in $\{-1, 0, 1\}^4 \setminus \{0\}$; the all-real argument remains the proof in Theorem 6. It also verifies the leading-ratio limit in (12), the endpoint-mixture identity, the variance identity, and the $\varepsilon = 1/4$ numerical threshold.

Analytic figure. Figure 1 is generated from $4\mu - 2\mu^2$ and $4\mu + 8\mu^2$ on rational points in $(0, 1/2]$. The supplement stores the rational numerators and denominators of these data. Neither curve comes from training or from a numerical minimization of curvature. In particular, the upper curve is only the explicit certificate proved for the two-parameter construction.