TRAINING A 2-TOKEN ATTENTION IS NP-HARD

Anonymous authors
Paper under double-blind review

ABSTRACT

As the title indicates, this work is inspired by the celebrated paper of Blum and Rivest, "Training a 3-node neural network is NP-complete." They showed that for a 2-layer, 3-node, n-input neural network (arguably the simplest neural network), deciding whether weights exist that fit a given training set exactly is NP-complete. We show that for an attention head with 1-dimensional query, key, and value, and n-dimensional token input (arguably the simplest attention head), the analogous exact-fit problem on a training set of 2-token sequences is NP-hard, and that if softmax is replaced with hardmax, the decision problem becomes NP-complete. These hardness results persist when fitting allows a fixed positive tolerance, measured coordinatewise or by mean squared loss.

1 INTRODUCTION

One of the best-known classical results in neural networks is that of Blum and Rivest (1988), establishing that training even a 3-node neural network is an NP-complete problem. The staying power of this result is impressive: Nearly four decades later, and limiting to the NeurIPS conference series alone, we see follow-up works nearly every year (Daniely & Vardi, 2020; Abrahamsen et al., 2021; Chen et al., 2022; Bertschinger et al., 2023; Froese & Hertrich, 2023; Kiani et al., 2024). This raises the corresponding question for the attention head (Vaswani et al., 2017): how simple can an attention head be before training becomes tractable?

A surprising aspect of Blum and Rivest's intractability result is the simplicity of the neural network they considered, just 2-layer, 3-node. We show that even when the dimensions of query, key, and value are lowered to the point of near-triviality, namely just 1-dimensional, i.e., the attention module associates each token with a scalar-valued key, query, and value (as opposed to vector-valued ones), and when training examples are just 2-token, again the lowest possible nontrivial value, the problem of training an attention head is already NP-hard. As in Blum and Rivest's construction, input dimension and sample count grow, while our query, key, and value dimensions may all remain one.

In the softmax construction, the same unknown maps determine token values and attention weights across all examples. Linear relations among the input tokens constrain sums of values and score differences. Exponentiation turns sums of score differences into products of ratios of attention weights. Together with the prescribed outputs, these relations force any exact fit to encode a satisfying Boolean assignment.

These hardness results persist under approximation. There is a fixed positive tolerance such that distinguishing datasets admitting an exact fit from those that cannot be fitted within that tolerance is NP-hard for both softmax and hardmax (Theorem 5). The same two-token examples suffice, with constant gaps for maximum coordinate error and total squared loss. We also obtain a constant mean squared loss gap with uniformly bounded inputs and targets (Theorem 6).

1.1 RELATED WORK

As we have alluded to, there has been an abundance of work following Blum & Rivest (1988) on the computational complexity of neural-network training and learning, including hardness results for small mean-squared-error objectives (Vu, 1998), approximation problems and richer architectures (Bartlett & Ben-David, 2002; Livni et al., 2014; Chen et al., 2022; Daniely & Vardi, 2020; Kiani et al., 2024), fixed input dimension and growing number of hidden neurons (Froese & Hertrich, 2023), other complexity classes like $\exists\mathbb{R}$-completeness (Abrahamsen et al., 2021; Bertschinger

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.

et al., 2023), and positive algorithmic results for special ReLU settings (Goel et al., 2017). Their common goal is to sharpen the contrast between the apparent simplicity of neural network architectures and the worst-case complexity of fitting their parameters.

Many studies address the expressivity, language-recognition power, and computational limits of transformers (Barceló et al., 2024; Bergsträßer et al., 2024; Alman & Yu, 2025; Bergsträßer et al., 2026). More directly related to fitting, Likhosherstov et al. (2023) show that weak approximation of a target self-attention matrix is NP-hard at any fixed key dimension greater than one, but polynomial-time solvable at key dimension one. Their weak approximation specifies an ordering of attention weights, and permits jointly choosing the input X and parameters $W_{\text{Q}}, W_{\text{K}}$. Ding et al. (2025) study single-layer linear attention. Given a minimizer in an enlarged parameter space, they show that recovering attention parameters realizing that minimizer is NP-hard. In concurrent work, Schachter (2026) reduces the existential theory of the reals with base-two exponentiation to zero-loss transformer training with three-token sequences and a growing number of heads. Value maps and output projections are fixed, and only selected outputs are supervised.

Our work on the other hand considers only a single attention head with scalar-valued queries, keys, and values, trained with mere two-token examples. With one-token examples, fitting becomes tractable (Proposition 1). Like Blum and Rivest (1988), we fix the input X as part of the training data and seek weights $W_{\text{Q}}, W_{\text{K}}, W_{\text{V}}$ fitting the outputs through an unknown value map.

Positive learning results also depend on the learner's information. Bhattamishra et al. (2026) recover the score matrix and nonzero value map of a single-head softmax regressor using $O(n^2)$ value queries at chosen one- and two-token inputs. Chen and Li (2025) give polynomial-time learning guarantees for a fixed number of heads under uniform Boolean inputs and parameter nondegeneracy assumptions. We ask whether arbitrary fixed rational data admit an exact or approximate fit.

1.2 NP-HARD VERSUS NP-COMPLETE

For softmax, we establish NP-hardness, while membership in NP remains open. To establish membership, one would need polynomial-length evidence of an exact fit that can be verified in polynomial time. Real fitting parameters, which may involve transcendental numbers, do not immediately provide such evidence. Section 3.4 gives general upper bounds for exact fitting and proves decidability under Schanuel's conjecture over the reals. Blum and Rivest faced no such difficulty: their network uses the sign activation. Replacing softmax by its zero-temperature limit, hardmax, gives NP-completeness at key dimension one (Theorem 4).

2 MODEL AND EXACT-FIT DECISION PROBLEMS

We define single-head attention, the two exact-fit problems, and the source 1-in-3-SAT variant.

2.1 SINGLE-HEAD ATTENTION LAYER

Throughout the paper we adopt the column-major convention for inputs: each sample is a matrix $X \in \mathbb{R}^{n\times p}$ whose p columns are n-dimensional input vectors. In large language models, a column vector in X is usually called a token. A *single-head attention layer* (Vaswani et al., 2017) with parameters

$$W_{\text{Q}} \in \mathbb{R}^{d_{\text{K}}\times n}, \qquad W_{\text{K}} \in \mathbb{R}^{d_{\text{K}}\times n}, \qquad W_{\text{V}} \in \mathbb{R}^{d_{\text{V}}\times n},$$

collectively denoted $\Theta = (W_{\text{Q}}, W_{\text{K}}, W_{\text{V}})$, maps $X \in \mathbb{R}^{n\times p}$ to $\mathbb{R}^{d_{\text{V}}\times p}$ via

$$A_{\text{S}}(X;\Theta) \;=\; W_{\text{V}}X\ \mathrm{softmax}\big(X^{\top}W_{\text{Q}}^{\top}W_{\text{K}}X\big) \tag{1}$$

in the softmax case, and via

$$A_{\text{H}}(X;\Theta) \;=\; W_{\text{V}}X\ \mathrm{hardmax}\big(X^{\top}W_{\text{Q}}^{\top}W_{\text{K}}X\big) \tag{2}$$

in the hardmax case. $W_{\text{Q}}X, W_{\text{K}}X, W_{\text{V}}X$ compute the query, key, and value, respectively. Both $\mathrm{softmax}$ and $\mathrm{hardmax}$ are applied *column-wise* to the $p \times p$ score matrix $X^{\top}W_{\text{Q}}^{\top}W_{\text{K}}X$. For a vector $z \in \mathbb{R}^p$,

$$\mathrm{softmax}(z)_i \;=\; \frac{e^{z_i}}{e^{z_1}+\cdots+e^{z_p}}, \qquad \mathrm{hardmax}(z) \;=\; \lim_{\tau\to+\infty}\mathrm{softmax}(\tau\, z),$$

so $\mathrm{hardmax}$ assigns weight $1/k$ to each of the k argmax coordinates and 0 elsewhere, the tie-averaging rule called *average hard attention* in the formal-language transformer literature (Barceló et al., 2024; Bergsträßer et al., 2024). (Alternatively, one can define a tie-breaking rule that assigns weight 1 to the first argmax coordinate and 0 elsewhere. It is not hard to see our argument still works for this tie-breaking $\mathrm{hardmax}$ *mutatis mutandis*.)

We write $M = W_{\mathrm{Q}}^{\top} W_{\mathrm{K}} \in \mathbb{R}^{n \times n}$ for the bilinear score matrix, and for two columns $a, b \in \mathbb{R}^n$,

$$s(a, b) \;=\; a^{\top} M b \;=\; (W_{\mathrm{Q}} a)^{\top} (W_{\mathrm{K}} b).$$

We write $v(a) = W_{\mathrm{V}} a \in \mathbb{R}^{d_{\mathrm{v}}}$ for the value of a. Concretely, the jth output column of (1) is

$$A_{\mathrm{S}}(X; \Theta)_{\bullet,\, j} \;=\; \sum_{i=1}^{p} v(x_i)\, \pi_{i,j}, \qquad \pi_{\bullet,\, j} \;=\; \mathrm{softmax}\big((s(x_i, x_j))_{i=1}^{p}\big), \tag{3}$$

and analogously for hardmax. Our hardness reductions construct a rank-one matrix $M = qk^{\top}$ for vectors $q, k \in \mathbb{R}^n$.

2.2 Exact-fit decision problems

Both decision problems below take a finite rational dataset and ask whether real parameters exist that fit it exactly; they differ only in the attention variant (softmax versus hardmax). We take $n, p_t \geq 1$; an empty dataset always has an exact fit.

Problem (SOFTMAX-EXACT-FIT)**.**

Input: *A finite dataset $\mathcal{D} = \{(X_t, Y_t) : t = 1, \ldots, N\}$[1] where $X_t \in \mathbb{Q}^{n \times p_t}$ and $Y_t \in \mathbb{Q}^{d_{\mathrm{v}} \times p_t}$.*

Output: *Decide whether there exist real parameters $W_{\mathrm{Q}}, W_{\mathrm{K}} \in \mathbb{R}^{d_{\mathrm{k}} \times n}$ and $W_{\mathrm{V}} \in \mathbb{R}^{d_{\mathrm{v}} \times n}$ such that $A_{\mathrm{S}}(X_t; \Theta) = Y_t$ for every $t = 1, \ldots, N$.*

Problem (HARDMAX-EXACT-FIT)**.**

Input: *The same finite dataset as in* SOFTMAX-EXACT-FIT.

Output: *Decide whether there exist real parameters Θ such that $A_{\mathrm{H}}(X_t; \Theta) = Y_t$ for every $t = 1, \ldots, N$.*

2.3 Length-one tractability

Before stating our hardness results we note that, with sequence length restricted to one, both decision problems collapse to rational linear feasibility.

Proposition 1. *The restrictions of* SOFTMAX-EXACT-FIT *and* HARDMAX-EXACT-FIT *to instances with $p_t = 1$ for every $t \in \{1, \ldots, N\}$ are decidable in polynomial time.*

Proof. A length-one input has a single column, so both $\mathrm{softmax}$ and $\mathrm{hardmax}$ assign weight 1 and $A_{\mathrm{S}}([x_t]; \Theta) = A_{\mathrm{H}}([x_t]; \Theta) = W_{\mathrm{V}} x_t$, independent of $(W_{\mathrm{Q}}, W_{\mathrm{K}})$. The exact-fit conditions thus reduce to the rational linear system $W_{\mathrm{V}} x_t = y_t$, decidable in polynomial time. □

Thus the reductions below must involve length-two samples; the remaining preliminary is the restricted 1-in-3-SAT source problem.

[1] We index training samples by t throughout; the indices i and j are reserved for the SAT variable and clause indices in our reductions (Sections 3–4).

2.4 SOURCE PROBLEM: DISTINCT-LITERAL 1-IN-3-SAT

Our reductions are from a slight restriction of the classical NP-complete 1-IN-3-SAT.

Problem (1-IN-3-SAT)**.**

Input: *A collection of clauses $C = \{c_1, \dots, c_m\}$ on Boolean variables $x_1, \dots, x_d$, where each clause $c_j = (\ell_{j,1}, \ell_{j,2}, \ell_{j,3})$ has three literals (each ℓ is some x_i or $\neg x_i$).*

Output: *Decide whether there is a truth assignment such that every clause has* exactly one *true literal.*

The problem is NP-complete by Schaefer (1978). The reductions of Sections 3 and 4 require the slight strengthening that clauses contain three *pairwise distinct* literals.

Lemma 1 (Distinct-literal restriction)**.** 1-IN-3-SAT *remains NP-complete when restricted to instances in which every clause has three pairwise distinct literals.*

Repeated literals can be eliminated by introducing new variables: a clause with two copies of a literal forces that literal to be false, and a clause with three copies is unsatisfiable. The full polynomial-time rewrite is given in Appendix A.

Sections 3 and 4 reduce from this distinct-literal restriction to the softmax and hardmax exact-fit problems respectively.

3 SOFTMAX EXACT-FIT IS NP-HARD

We begin with the softmax case, reducing from the distinct-literal restriction of 1-in-3-SAT (Lemma 1).

Theorem 1 (Softmax attention is NP-hard)**.** *For every fixed $d_{\mathrm{K}} \geq 1$ and $d_{\mathrm{V}} \geq 1$,* SOFTMAX-EXACT-FIT *is NP-hard. Hardness already holds when every input sequence has length $p_t \leq 2$ and every entry of every X_t lies in $\{-2, -1, 0, 1\}$.*

We first prove the theorem for $d_{\mathrm{K}} = d_{\mathrm{V}} = 1$; Section 4.2 gives the general case.

3.1 THE REDUCTION

Given a distinct-literal 1-IN-3-SAT instance ϕ with d variables $x_1, \dots, x_d$ and m clauses $c_j = (\ell_{j,1}, \ell_{j,2}, \ell_{j,3})$ (Lemma 1), set $n = 2d + 1$ and use the coordinate basis vectors

$$F, \;\; L_1, \dots, L_d, \;\; \bar{L}_1, \dots, \bar{L}_d \;\in\; \{0, 1\}^n.$$

For a literal ℓ, write $T(\ell) = L_i$ if $\ell = x_i$ and $T(\ell) = \bar{L}_i$ if $\ell = \neg x_i$. This means we will encode the literals x_i and $\neg x_i$ as vectors $L_i, \bar{L}_i$.

For each variable i and clause j, define the vectors

$$P_i \;=\; L_i + \bar{L}_i - F, \qquad D_j \;=\; T(\ell_{j,1}) + T(\ell_{j,2}) + T(\ell_{j,3}) - 2F.$$

Because each clause has three pairwise distinct literals (Lemma 1), the three terms in D_j are distinct basis vectors, so D_j has coordinate 1 at three literal positions, -2 at F, and 0 elsewhere. Every input vector used in the reduction therefore has integer entries in $\{-2, -1, 0, 1\}$.

The dataset $\mathcal{D}_\phi$ consists of $4d + m + 1$ samples in six groups, with output dimension $d_{\mathrm{V}} = 1$:

Group	Sample X_t	Target Y_t	Count
1	$[F]$	0	1
2	$[P_i]$	5	d
3	$[F,\, L_i]$	$(1,\, 1)$	d
4	$[F,\, \bar{L}_i]$	$(1,\, 1)$	d
5	$[F,\, P_i]$	$(5/3,\, 5/2)$	d
6	$[F,\, D_j]$	$(7/3,\, 14/3)$	m

The first-column equations for Groups 3–6 are easiest to read as equations in softmax odds, which we introduce next.

Define, for each variable x_i, the strictly positive scalars

$$\alpha_i = e^{s(F,F)-s(L_i,F)}, \quad \beta_i = e^{s(F,F)-s(\bar{L}_i,F)}, \quad \gamma_i = e^{s(F,F)-s(P_i,F)}, \tag{4}$$

which by bilinearity of s and $P_i = L_i + \bar{L}_i - F$ satisfy the exact identity $\gamma_i = \alpha_i\beta_i$. Groups 1–5 force $\alpha_i + \beta_i = 3$ and $\alpha_i\beta_i = 2$ (Lemma 2), pinning $\{\alpha_i, \beta_i\} = \{1, 2\}$ and yielding the Boolean assignment $x_i = \text{TRUE}$ iff $\alpha_i = 2$. Group 6 then forces exactly one true literal per clause (Lemma 3). All soundness arguments use only the first entry of Y_t; the second entries are recovered automatically by the explicit completeness construction (Lemma 4).

3.2 Soundness

Soundness proceeds in two steps: the Variable Lemma extracts a Boolean assignment from the fitted parameters, and the Clause Lemma checks the 1-in-3 condition on each clause.

Lemma 2 (Variable Lemma). *If parameters Θ exactly fit Groups 1–5 of $\mathcal{D}_\phi$, then for every $i \in \{1, \ldots, d\}$,*

$$\alpha_i + \beta_i \;=\; 3, \qquad \alpha_i\beta_i \;=\; 2,$$

and consequently $\{\alpha_i, \beta_i\} = \{1, 2\}$ and $\{v(L_i), v(\bar{L}_i)\} = \{2, 3\}$.

Proof. The equation $A_{\text{s}}(X; \Theta) = Y$ for length-one samples $[F] \mapsto 0$ and $[P_i] \mapsto 5$ reduces to $v(F) = 0$ and $v(P_i) = 5$. For the length-two samples $[F, L_i]$, $[F, \bar{L}_i]$, $[F, P_i]$, the first entry of the same equation yields

$$\frac{v(L_i)}{1+\alpha_i} = 1, \qquad \frac{v(\bar{L}_i)}{1+\beta_i} = 1, \qquad \frac{v(P_i)}{1+\gamma_i} = \tfrac{5}{3}.$$

Hence $v(L_i) = 1+\alpha_i$, $v(\bar{L}_i) = 1+\beta_i$, and $\gamma_i = 2$. Linearity of W_{v} together with $P_i = L_i + \bar{L}_i - F$ gives $v(L_i) + v(\bar{L}_i) = 5$, so $\alpha_i + \beta_i = 3$; bilinearity gives $\gamma_i = \alpha_i\beta_i$, hence $\alpha_i\beta_i = \gamma_i = 2$. The strictly positive scalars α_i, β_i are therefore the two roots of $z^2 - 3z + 2 = (z-1)(z-2)$, so $\{\alpha_i, \beta_i\} = \{1, 2\}$ and $\{v(L_i), v(\bar{L}_i)\} = \{2, 3\}$. □

Define the Boolean assignment

$$x_i = \text{TRUE} \iff \alpha_i = 2 \iff v(L_i) = 3,$$

so a literal ℓ is true iff $\rho(\ell) = 2$, where $\rho(x_i) = \alpha_i$ and $\rho(\neg x_i) = \beta_i$.

Lemma 3 (Clause Lemma). *If Θ exactly fits Groups 1–6, then under the Boolean assignment above, every clause c_j has exactly one true literal.*

Proof. By bilinearity, with $\delta_j = e^{s(F,F)-s(D_j,F)}$,

$$s(F,F) - s(D_j,F) \;=\; \sum_{r=1}^{3} \big(s(F,F) - s(T(\ell_{j,r}),F)\big), \quad \text{hence} \quad \delta_j \;=\; \prod_{r=1}^{3} \rho(\ell_{j,r}).$$

Let k_j be the number of true literals in c_j. Then $\delta_j = 2^{k_j}$ and, since $v(T(\ell)) \in \{2, 3\}$ with $v = 3$ on true literals,

$$v(D_j) \;=\; \sum_{r=1}^{3} v(T(\ell_{j,r})) - 2\,v(F) \;=\; 3k_j + 2(3 - k_j) \;=\; 6 + k_j.$$

The equation for the first entry of Y_t in Group 6 is therefore

$$\frac{v(D_j)}{1+\delta_j} \;=\; \frac{6+k_j}{1+2^{k_j}}.$$

Setting equal to $7/3$ leaves only $k_j = 1$:

k_j	0	1	2	3
$(6+k_j)/(1+2^{k_j})$	3	$\mathbf{7/3}$	$8/5$	1

□

3.3 COMPLETENESS

Completeness is by explicit construction: we exhibit row-vector parameters $(W_{\mathrm{Q}}, W_{\mathrm{K}}, W_{\mathrm{V}})$ derived from a satisfying assignment of ϕ and verify the targets group by group.

Lemma 4 (Completeness)**.** *If ϕ is satisfiable, then there exist parameters Θ with $d_{\mathrm{K}} = d_{\mathrm{V}} = 1$ that exactly fit $\mathcal{D}_\phi$.*

Proof. Fix a satisfying assignment for ϕ. With $d_{\mathrm{K}} = d_{\mathrm{V}} = 1$ the parameters $W_{\mathrm{Q}}, W_{\mathrm{K}}, W_{\mathrm{V}} \in \mathbb{R}^{1\times n}$ are row vectors. Define them on the basis vectors by

Vectors	W_{V}	W_{Q}	W_{K}
F	0	0	1
True literal $T(\ell)$	3	$-\log 2$	1
False literal $T(\ell)$	2	0	0

extended linearly to all of $\mathbb{R}^n$. Then $v(F) = 0$ (Group 1) and, since exactly one of $L_i, \bar{L}_i$ is true, $v(P_i) = 3 + 2 = 5$ (Group 2). For Groups 3–6, the score matrix of $[F, U]$ has columnwise entries $(0,\ W_{\mathrm{Q}}U)^\top$ and $(0,\ (W_{\mathrm{Q}}U)(W_{\mathrm{K}}U))^\top$ (using $W_{\mathrm{Q}}F = 0, W_{\mathrm{K}}F = 1$). Direct substitution gives $(1, 1)$ for literal samples, $(5/3, 5/2)$ for P_i, and $(7/3, 14/3)$ for D_j. Appendix B.1 gives the four cases and the clause calculation. □

The constructed dataset has $N = 4d + m + 1$ samples in dimension $n = 2d + 1$. This completes the polynomial-time many-one reduction proving Theorem 1 in the case of $d_{\mathrm{K}} = d_{\mathrm{V}} = 1, p_t \le 2$.

3.4 UPPER BOUNDS FOR SOFTMAX EXACT FITTING

Let $\mathbb{R}_{\exp} = (\mathbb{R}, +, \cdot, <, 0, 1, \exp)$, and let $\mathrm{ETh}^{\exists}(\mathbb{R}_{\exp})$ denote the truth problem for existential first-order sentences over this structure; the class of problems polynomial-time reducible to it is denoted $\exists\mathbb{R}_{\exp}$. This is the natural setting for exact softmax equations: it retains a finite Turing encoding while allowing the exponential function explicitly.

Theorem 2 (Upper bounds for softmax exact fitting)**.** *For every fixed $d_{\mathrm{K}}, d_{\mathrm{V}} \ge 1$,* SOFTMAX-EXACT-FIT *has a polynomial-time many-one reduction to $\mathrm{ETh}^{\exists}(\mathbb{R}_{\exp})$ and, unconditionally, belongs to Σ_2^0. If Schanuel's conjecture for real numbers holds, then* SOFTMAX-EXACT-FIT *is decidable.*

The exponential encoding is direct; see also Schachter (2026, Proposition 12) for a general transformer encoding. For each sample $t = 1, \ldots, N$, write the columns of X_t as $x_{t,1}, \ldots, x_{t,p_t}$, and introduce variables z_{tab}, u_{tab} for all token positions $a, b = 1, \ldots, p_t$. Along with variables for the entries of $W_{\mathrm{Q}}, W_{\mathrm{K}}, W_{\mathrm{V}}$, impose (5) for all such a, b, and (6) for all output rows $c = 1, \ldots, d_{\mathrm{V}}$ and token positions $b = 1, \ldots, p_t$:

$$z_{tab} = (W_{\mathrm{Q}}x_{t,a})^\top (W_{\mathrm{K}}x_{t,b}), \qquad u_{tab} = \exp(z_{tab}), \tag{5}$$

$$\sum_{a=1}^{p_t} \big((W_{\mathrm{V}}x_{t,a})_c - (Y_t)_{cb}\big) u_{tab} = 0. \tag{6}$$

Since every u_{tab} is positive, equation (6) is exactly the softmax fit equation after multiplication by its denominator. The number and size of these equations are polynomial in the explicitly encoded dataset; rational denominators can be cleared and binary integer constants can be constructed by repeated-doubling constraints. Their existential closure is therefore the claimed polynomial-size sentence.

For the unconditional bound, substitute the score exponentials into (6) and sum the squared residuals. The minimum on each parameter box with integer bounds is computable. Exact fitting asks whether some box has minimum zero, a condition expressible by computable tests at every approximation precision. This existential choice followed by universal testing gives the Σ_2^0 bound (Appendix C), without an unconditional decision procedure.

Under Schanuel's conjecture over the reals, Macintyre & Wilkie (1996) prove decidability of the complete theory of $\mathbb{R}_{\exp}$, which in particular decides our existential sentences.

4 HARDMAX EXACT-FIT IS NP-HARD AND NP-COMPLETE AT $d_{\mathrm{K}} = 1$

We next turn to hardmax attention. NP-hardness holds for all fixed key and value dimensions (Theorem 3). When the key dimension is one, an additional NP-membership argument upgrades this to NP-completeness (Theorem 4).

Theorem 3 (Hardmax attention is NP-hard)**.** *For every fixed $d_{\mathrm{K}} \geq 1$ and $d_{\mathrm{V}} \geq 1$,* HARDMAX-EXACT-FIT *is NP-hard. Hardness already holds when every input sequence has length $p_t \leq 2$ and every entry of every X_t is Boolean (in $\{0, 1\}$).*

Theorem 4 (Hardmax attention is NP-complete)**.** *For every fixed $d_{\mathrm{V}} \geq 1$ and $d_{\mathrm{K}} = 1$,* HARDMAX-EXACT-FIT *is in NP, and hence (by Theorem 3 at $d_{\mathrm{K}} = 1$) is NP-complete.*

4.1 PROOF OF THEOREM 3

The hardmax reduction keeps the Boolean core of the softmax construction but removes the auxiliary vector F: truth is encoded directly by values 0 and 1. Given a distinct-literal 1-IN-3-SAT instance ϕ on d variables $x_1, \ldots, x_d$ and m clauses c_j (Lemma 1), set $n = 2d$ and use the coordinate basis vectors

$$L_1, \ldots, L_d, \;\; \bar{L}_1, \ldots, \bar{L}_d \;\in\; \{0, 1\}^n.$$

For each variable i and clause j, define the vectors

$$P_i \;=\; L_i + \bar{L}_i, \qquad D_j \;=\; T(\ell_{j,1}) + T(\ell_{j,2}) + T(\ell_{j,3}).$$

By Lemma 1, D_j is a sum of three distinct basis vectors. Every input vector therefore has Boolean entries in $\{0, 1\}$.

The dataset $\mathcal{D}_\phi$ consists of $2d + m$ samples in three groups, with $d_{\mathrm{V}} = 1$:

Group	Sample X_t	Target Y_t	Count
1	$[P_i]$	1	d
2	$[L_i,\, \bar{L}_i]$	$(1,\, 0)$	d
3	$[D_j]$	1	m

Soundness. Suppose Θ exactly fits $\mathcal{D}_\phi$.

Group 1. The sample $[P_i] \mapsto 1$ has trivial hardmax weight 1, so $v(P_i) = 1$. By linearity of W_{V},

$$v(L_i) + v(\bar{L}_i) \;=\; 1. \tag{7}$$

Group 2. For the length-two sample $[L_i, \bar{L}_i] \mapsto (1, 0)$, each of the two output columns equals a $\mathrm{hardmax}$-weighted average of $v(L_i)$ and $v(\bar{L}_i)$. If column j of the score matrix has both entries equal, the equal-weight tiebreak rule (Section 2) assigns weight $1/2$ to each, giving column-j output $\frac{1}{2}(v(L_i) + v(\bar{L}_i)) = \frac{1}{2}$ by (7), contradicting both targets 1 and 0. Hence both columns have a strict winner: column 1 selects either L_i or $\bar{L}_i$, forcing the selected value to equal 1; column 2 analogously forces the selected value to equal 0. Together with (7),

$$\{v(L_i),\, v(\bar{L}_i)\} \;=\; \{0,\, 1\}.$$

Define the Boolean assignment $x_i = \textsc{true}$ iff $v(L_i) = 1$, so a literal ℓ is true iff $v(T(\ell)) = 1$.

Group 3. The length-1 sample $[D_j] \mapsto 1$ gives $v(D_j) = \sum_{r=1}^{3} v(T(\ell_{j,r})) = 1$. Each summand is in $\{0, 1\}$, so exactly one of the three literals of c_j is true.

Completeness. If ϕ is satisfiable, fix a satisfying assignment and define row vectors $W_{\mathrm{V}}, W_{\mathrm{Q}} \in \mathbb{R}^{1\times n}$ by

$$W_{\mathrm{V}}T(\ell) \;=\; W_{\mathrm{Q}}T(\ell) \;=\; \begin{cases} 1, & \ell \text{ is true,} \\ 0, & \ell \text{ is false,} \end{cases}$$

and $W_{\mathrm{K}} \in \mathbb{R}^{1\times n}$ by $W_{\mathrm{K}}L_i = 1$, $W_{\mathrm{K}}\bar{L}_i = -1$ for every i (independent of truth). Group 1 holds because exactly one of $L_i, \bar{L}_i$ is true, so $v(P_i) = 1 + 0 = 1$. Group 3 holds because each clause has exactly one true literal, so $v(D_j) = 1 + 0 + 0 = 1$.

For Group 2, the displayed parameters select the true literal in the first column and the false literal in the second, giving the target $(1, 0)$. Appendix B.2 gives the score calculation.

4.2 The $d_{\mathrm{K}} > 1$ and $d_{\mathrm{V}} > 1$ case

For $d_{\mathrm{K}} > 1$, add zero rows to the scalar completeness parameters $W_{\mathrm{Q}}, W_{\mathrm{K}}$. This preserves $M = W_{\mathrm{Q}}^{\top}W_{\mathrm{K}}$. Soundness holds for arbitrary $M \in \mathbb{R}^{n\times n}$. The softmax argument uses only bilinearity through $\gamma_i = \alpha_i\beta_i$ and $\delta_j = \prod_r \rho(\ell_{j,r})$. The hardmax argument uses only that a two-token weight column is $(1, 0)^{\top}$, $(0, 1)^{\top}$, or $(1/2, 1/2)^{\top}$.

For $d_{\mathrm{V}} > 1$, replicate each scalar target and the completeness row W_{V}. Conversely, the first row of any fitted value matrix fits the scalar targets with the same $W_{\mathrm{Q}}, W_{\mathrm{K}}$, since attention weights do not depend on W_{V}. This proves Theorems 1 and 3 for all fixed positive $d_{\mathrm{K}}, d_{\mathrm{V}}$.

4.3 NP-completeness at $d_{\mathrm{K}} = 1$ (proof of Theorem 4)

Theorem 3 at $d_{\mathrm{K}} = 1$ already gives NP-hardness. To obtain NP-completeness it suffices to show NP-membership, which we do by exhibiting a polynomial-bit-length rational certificate.

Lemma 5 (Rational certificate for linear feasibility)**.** *A consistent finite system of rational linear equations and weak/strict inequalities in ν real variables, whose constraint coefficients have total bit length B, has a rational solution of bit length polynomial in $\nu + B$.*

This is standard rational linear-programming bit-complexity; mixed weak/strict rational systems are handled by the well-known reduction to standard LP feasibility (Schrijver, 1986).

At $d_{\mathrm{K}} = 1$, each score column is a scalar multiple of the same query vector. Its maximizing set is therefore the query argmax, the query argmin, or all positions, according to the sign of the scalar key. Guessing these signs and order relations leaves rational linear constraints on $W_{\mathrm{Q}}, W_{\mathrm{K}}$, and the prescribed outputs give rational linear equations in W_{V}. Lemma 5 supplies polynomial-length rational witnesses. Appendix D gives the full certificate and verifier.

The argument relies on $d_{\mathrm{K}} = 1$ for the rank-one factorization. For $d_{\mathrm{K}} \geq 2$, this certificate argument does not establish NP membership.

5 Approximate fitting remains NP-hard

The preceding reductions also establish hardness of approximate fitting. On the same datasets, every choice of parameters leaves a fixed positive total squared error whenever the formula is unsatisfiable. For $\tau \in \{\mathrm{S}, \mathrm{H}\}$, write

$$R_t^{\tau}(\Theta) = Y_t - A_{\tau}(X_t; \Theta), \qquad L_{\tau}(\Theta; \mathcal{D}) = \sum_{t=1}^{N} \|R_t^{\tau}(\Theta)\|_F^2,$$

$$E_{\tau}(\Theta; \mathcal{D}) = \max_{t,r,c} |(R_t^{\tau}(\Theta))_{r,c}|, \qquad \mathrm{OPT}_{\tau}(\mathcal{D}) = \inf_{\Theta} L_{\tau}(\Theta; \mathcal{D}).$$

Theorem 5 (Approximate fitting remains NP-hard)**.** *For every fixed $d_{\mathrm{K}}, d_{\mathrm{V}} \geq 1$, each of the following promise problems is NP-hard:*

$$\textit{softmax:} \quad \mathrm{OPT}_{\mathrm{S}}(\mathcal{D}) = 0 \quad \textit{versus} \; \mathrm{OPT}_{\mathrm{S}}(\mathcal{D}) \geq 10^{-6},$$

$$\textit{hardmax:} \quad \mathrm{OPT}_{\mathrm{H}}(\mathcal{D}) = 0 \quad \textit{versus} \; \mathrm{OPT}_{\mathrm{H}}(\mathcal{D}) \geq \frac{1}{5}.$$

Zero loss is attained on the constructed yes-instances. The reductions are exactly those of Sections 3 and 4; in particular, $p_t \le 2$, and the input entries lie in $\{-2,-1,0,1\}$ for softmax and in $\{0,1\}$ for hardmax.

The same reductions give coordinatewise separation. In every softmax no-instance, every Θ satisfies $E_{\mathrm{S}}(\Theta;\mathcal{D}) > 10^{-3}$; in every hardmax no-instance, every Θ satisfies $E_{\mathrm{H}}(\Theta;\mathcal{D}) \ge 1/4$.

For these datasets, averaging the total loss over the N samples gives the guaranteed gaps $10^{-6}/N$ and $1/(5N)$. The exact and approximate hardness results also hold with exactly two tokens per sample and uniformly bounded parameter entries (Appendix E.4).

Completeness is unchanged. For softmax soundness, if every coordinate error is at most $\eta \le 10^{-3}$, the variable and clause samples give

$$|\alpha_i + \beta_i - 3| \le 12\eta, \quad |\alpha_i\beta_i - 2| \le 6\eta, \quad |3(6+k) - 7(2^k+1)| \le 3396\eta.$$

The first two inequalities place the two positive odds within 32η of 1 and 2, one near each. Declare x_i true when α_i is near 2, and false otherwise. The last inequality forces the number k of true literals in each clause to be one: at $\eta = 10^{-3}$ its right side is 3.396, while its left side is $4, 0, 11, 36$ for $k = 0, 1, 2, 3$. Since $E_{\mathrm{S}} \le \sqrt{L_{\mathrm{S}}}$, this also proves the total-loss gap.

For hardmax, a tied variable column and the sample $[P_i] \mapsto 1$ contribute at least $1/5$ to total loss, while two columns with the same strict winner contribute at least $1/2$. Below $1/5$, declare the first column's winning literal true. The three literal residuals and the clause residual occupy four distinct coordinates, giving $|k-1| \le 2\sqrt{L_{\mathrm{H}}} < 1$ by Cauchy–Schwarz. A separate coordinatewise argument gives $1/4$.

Soundness holds for arbitrary M, hence every key dimension. For any value dimension, the replicated targets allow projection onto one output row. Appendix E gives the complete inequalities.

Theorem 6 (Mean squared loss remains NP-hard)**.** *There is a constant $c > 0$ such that, for every fixed $d_{\mathrm{K}}, d_{\mathrm{V}} \ge 1$ and each $\tau \in \{\mathrm{S}, \mathrm{H}\}$, it is NP-hard to distinguish nonempty datasets admitting an exact fit from those satisfying $\mathrm{OPT}_\tau(\mathcal{D})/N \ge c$. Every sample may have exactly two tokens, with the same bounded input and target alphabets as in Theorem 5. The yes-instances admit fitting parameters with entries in $[-3,3]$ for softmax and $[-1,1]$ for hardmax.*

We reduce from a version of 1-in-3-SAT in which every assignment violates a fixed fraction of clauses in the no-case. Each clause is supplied with its own copies of the samples used to recover its literal values. The shared parameters determine one Boolean assignment, and the local error estimates force a fixed loss for each violated clause. Appendix F gives the construction and proof.

Causal attention. For softmax, exact-fit and mean squared loss hardness also hold with a causal mask and only the last scalar output of each two-token sequence supervised. Proposition 2 in Appendix G gives the construction and proof, with the same uniform bounds.

6 Concluding remarks

Blum and Rivest (1988) showed that training even a simple neural network is hard. Our corresponding worst-case results for a single attention head cover exact fitting and fixed tolerances in coordinatewise error and mean squared loss, with uniformly bounded inputs and targets.

Satisfiable formulas give rank-one fitting parameters in both reductions, while the soundness arguments hold for an arbitrary score matrix M.

Beyond NP for softmax. Theorem 2 gives an unconditional Σ_2^0 upper bound and decidability under Schanuel's conjecture over the reals. Determining the precise complexity of scalar two-token softmax fitting remains open.

Polynomial-time algorithms. A complementary direction is to identify restricted data or attention architectures that admit efficient fitting algorithms.

Reproducibility statement

Section 2 specifies the attention model and fitting problems. Sections 3 and 4 give the reductions and fitting parameters explicitly. Appendices A–G provide the remaining proofs and construction details.

AI use statement

Generative AI tools were used in literature review, editing, and formatting during drafting and revisions of the manuscript. The mathematical content, including definitions, theorem statements, reductions, and proofs, originates with the authors, who take full responsibility for its correctness.

References

Mikkel Abrahamsen, Linda Kleist, and Tillmann Miltzow. Training neural networks is $\exists\mathbb{R}$-complete. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 18293–18306, 2021.

Josh Alman and Hantao Yu. Fundamental limitations on subquadratic alternatives to transformers. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Pablo Barceló, Alexander Kozachinskiy, Anthony Widjaja Lin, and Vladimir Podolskii. Logical languages accepted by transformer encoders with hard attention. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Peter L. Bartlett and Shai Ben-David. Hardness results for neural network approximation problems. *Theoretical Computer Science*, 284(1):53–66, 2002.

Pascal Bergsträßer, Chris Köcher, Anthony Widjaja Lin, and Georg Zetzsche. The power of hard attention transformers on data sequences: A formal language theoretic perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Pascal Bergsträßer, Ryan Cotterell, and Anthony Widjaja Lin. Transformers are inherently succinct. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Daniel Bertschinger, Christoph Hertrich, Paul Jungeblut, Tillmann Miltzow, and Simon Weber. Training fully connected neural networks is $\exists\mathbb{R}$-complete. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Satwik Bhattamishra, Kulin Shah, Michael Hahn, and Varun Kanade. Provably learning attention with queries. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2601.16873.

Avrim Blum and Ronald L. Rivest. Training a 3-node neural network is NP-complete. In *Advances in Neural Information Processing Systems (NIPS)*, volume 1, pp. 494–501, 1988.

Sitan Chen and Yuanzhi Li. Provably learning a multi-head attention layer. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 1744–1754, 2025. doi: 10.1145/3717823.3718174.

Sitan Chen, Aravind Gollakota, Adam Klivans, and Raghu Meka. Hardness of noise-free learning for two-hidden-layer neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 10709–10724, 2022.

Amit Daniely and Gal Vardi. Hardness of learning neural networks with natural weights. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 930–940, 2020.

Yanna Ding, Songtao Lu, Yingdong Lu, Tomasz J. Nowicki, and Jianxi Gao. Optimality and NP-hardness of transformers in learning Markovian dynamical functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. URL https://openreview.net/forum?id=ZKYxfU8de7.

Vincent Froese and Christoph Hertrich. Training neural networks is NP-hard in fixed dimension. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 44039–44049, 2023.

Surbhi Goel, Varun Kanade, Adam Klivans, and Justin Thaler. Reliably learning the ReLU in polynomial time. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1004–1042. PMLR, 2017.

Teemu Hankala, Miika Hannula, Juha Kontinen, and Jonni Virtema. Complexity of neural network training and ETR: Extensions with effectively continuous functions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11):12278–12285, 2024. doi: 10.1609/aaai.v38i11.29118.

Johan Håstad. Some optimal inapproximability results. *Journal of the ACM*, 48(4):798–859, 2001.

Bobak T. Kiani, Jason Wang, and Melanie Weber. Hardness of learning neural networks under the manifold hypothesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Valerii Likhosherstov, Krzysztof Choromanski, and Adrian Weller. On the expressive flexibility of self-attention matrices. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37(7), pp. 8773–8781, 2023. doi: 10.1609/aaai.v37i7.26055.

Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, pp. 855–863, 2014.

Angus Macintyre and A. J. Wilkie. On the decidability of the real exponential field. In Piergiorgio Odifreddi (ed.), *Kreiseliana: About and Around Georg Kreisel*, pp. 441–467. A K Peters, 1996.

Nicholas Schachter. Transformers speak base-2: Attention training and the existential theory of $(\mathbb{R}, 2^x)$. SSRN preprint, 2026. URL `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=7360439`.

Thomas J. Schaefer. The complexity of satisfiability problems. In *Proceedings of the Tenth Annual ACM Symposium on Theory of Computing (STOC)*, pp. 216–226, 1978.

Alexander Schrijver. *Theory of Linear and Integer Programming*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, 1986.

Lucio Tanzini. Existential theory of the reals, neural networks, and Tarski's exponential function problem. Master's thesis, Università di Pisa, 2025. URL `https://etd.adm.unipi.it/t/etd-04232025-100531/`. Public abstract; full text unavailable.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, pp. 5998–6008, 2017.

Van H. Vu. On the infeasibility of training neural networks with small mean-squared error. *IEEE Transactions on Information Theory*, 44(7):2892–2900, 1998.

A THE DISTINCT-LITERAL RESTRICTION

Proof of Lemma 1. Given an instance ϕ, scan its clauses and rewrite each non-distinct clause as follows.

Case 1. A clause of the form (ℓ, ℓ, ℓ) (all three literals equal) is unsatisfiable in 1-in-3 semantics, so ϕ itself is unsatisfiable. Output two distinct-literal clauses

$$(x, y, \neg y), \qquad (\neg x, z, \neg z)$$

on three new variables x, y, z. The first clause forces x to be false and the second forces x to be true, so the two new clauses are unsatisfiable.

Case 2. A clause of the form (ℓ, ℓ, ℓ') with $\ell \neq \ell'$ (after relabeling so the two equal literals are listed first) forces both copies of ℓ to take the same truth value; the 1-in-3 condition then forces ℓ to be false and ℓ' to be true. Replace this clause by the two distinct-literal clauses

$$(\ell, x, \neg x), \qquad (\neg \ell', x, \neg x)$$

with a new variable x. Each new clause has three pairwise distinct literals. The two clauses force ℓ to be false and ℓ' to be true.

Applying these replacements independently to all non-distinct clauses gives an equisatisfiable instance of polynomial size, and membership in NP is immediate. □

B VERIFICATION OF THE COMPLETENESS PARAMETERS

B.1 SOFTMAX

For the parameters in Lemma 4, Groups 1 and 2 follow from the displayed basis values. For Groups 3–6, the score columns of $[F, U]$ are $(0, W_{\mathrm{Q}}U)^\top$ and $(0, (W_{\mathrm{Q}}U)(W_{\mathrm{K}}U))^\top$. The four resulting cases are:

Sample $[F, U]$ (case)	$(W_{\mathrm{Q}}U,\, W_{\mathrm{K}}U)$	F-col output	U-col output
$U \in \{L_i, \bar{L}_i\}$, U true	$(-\log 2,\, 1)$	1	1
$U \in \{L_i, \bar{L}_i\}$, U false	$(0,\, 0)$	1	1
$U = P_i$ (one literal true)	$(-\log 2,\, 0)$	$5/3$	$5/2$
$U = D_j$ (1-in-3 satisfied)	$(-\log 2,\, -1)$	$7/3$	$14/3$

For example, the D_j case has score matrix $\left(\begin{smallmatrix} 0 & 0 \\ -\log 2 & \log 2 \end{smallmatrix}\right)$, columnwise softmax $(2/3, 1/3)^\top$ and $(1/3, 2/3)^\top$, and $v(D_j) = 7$ (one true plus two false literals); the outputs are $(0, 7) \cdot (2/3, 1/3)^\top = 7/3$ and $(0, 7) \cdot (1/3, 2/3)^\top = 14/3$. The other cases are verified similarly.

B.2 HARDMAX

For Group 2 when x_i is true, $W_{\mathrm{Q}}L_i = 1, W_{\mathrm{Q}}\bar{L}_i = 0$; the score matrix of $[L_i, \bar{L}_i]$ is $\left(\begin{smallmatrix} 1 & -1 \\ 0 & 0 \end{smallmatrix}\right)$. Column 1 has unique max in row 1 (since $1 > 0$), giving weights $(1, 0)^\top$ and output $v(L_i) = 1$. Column 2 has unique max in row 2 (since $0 > -1$), giving weights $(0, 1)^\top$ and output $v(\bar{L}_i) = 0$. The output is $(1, 0)$, matching the target. The case when x_i is false is similar.

C PROOF OF THE SOFTMAX UPPER BOUNDS

An empty dataset has an exact fit and may be mapped to the true sentence $0 = 0$. For a nonempty dataset, we first justify the polynomial encoding in Section 3.4. Equation (5) expands to a quadratic polynomial with rational coefficients in the parameter entries, and (6) is polynomial in those entries and the variables u_{tab}. Clearing rational denominators preserves the solution set. In the language with constants 0 and 1, an integer given in binary is constructed with a linear-size chain $c_0 = 0$ and $c_{i+1} = c_i + c_i + b_i$, where $b_i \in \{0, 1\}$ is its next bit; negative terms are moved across the equality.

Hence the construction is polynomial in the bit length of the dataset. Conversely, any solution of the existential sentence has $u_{tab} = e^{z_{tab}} > 0$, so division of (6) by $\sum_a u_{tab}$ recovers the original attention equation. This proves the many-one reduction.

We next prove the unconditional Σ_2^0 bound without invoking a normal form theorem for general exponential formulas. Let θ list the entries of $(W_{\mathrm{Q}}, W_{\mathrm{K}}, W_{\mathrm{V}})$, and define the cleared residuals

$$E_{tcb}(\theta) = \sum_{a=1}^{p_t} \big((W_{\mathrm{V}} x_{t,a})_c - (Y_t)_{cb}\big) \exp\big((W_{\mathrm{Q}} x_{t,a})^\top (W_{\mathrm{K}} x_{t,b})\big), \qquad G_I(\theta) = \sum_{t,c,b} E_{tcb}(\theta)^2. \quad (8)$$

Thus I is exactly fit if and only if G_I has a real zero. For an integer $R \geq 1$, compactness gives the attained minimum

$$\mu_{I,R} := \min_{\theta \in [-R,R]^D} G_I(\theta), \qquad D = (2d_{\mathrm{K}} + d_{\mathrm{V}})n.$$

This minimum is a computable real, uniformly in (I, R). Indeed, on the box all score polynomials and their derivatives have a rational bound computable from (I, R). For a polynomial, summing absolute coefficient bounds times powers of R bounds its value and each derivative. If $B \geq 0$ bounds a score, then $e^B \leq 3^{\lceil B \rceil}$. Applying these bounds after differentiating G_I gives rational upper bounds on the absolute values of its partial derivatives. Their sum is a computable rational L with $|G_I(\theta) - G_I(\theta')| \leq L\|\theta - \theta'\|_\infty$ on the box. At a rational point, the Taylor series with its standard remainder bound computes each exponential, and hence G_I, to any prescribed rational error. Given $\varepsilon > 0$, evaluate G_I with error at most $\varepsilon/3$ on a finite rational grid such that every point of the box is within $\varepsilon/(3L)$ in the maximum norm of a grid point (using one point if $L = 0$). The least computed value differs from $\mu_{I,R}$ by at most ε. This is a terminating, though not necessarily efficient, algorithm.

For a valid instance and $R \geq 1$, let $a_{I,R,k} \in \mathbb{Q}$ be the output of the fixed approximation algorithm above, with $|a_{I,R,k} - \mu_{I,R}| \leq 2^{-(k+3)}$, and define the decidable relation

$$\mathcal{R}(I, R, k) = \begin{cases} \text{false}, & I \text{ is malformed or } R < 1, \\ [a_{I,R,k} < 2^{-k}], & \text{otherwise}, \end{cases}$$

where brackets denote the truth value of the rational comparison. The approximation is computed only in the second case. For every $R \geq 1$,

$$\mu_{I,R} = 0 \quad \Longleftrightarrow \quad \forall k \in \mathbb{N}\ \mathcal{R}(I, R, k).$$

The forward implication follows from $a_{I,R,k} \leq 2^{-(k+3)} < 2^{-k}$. If $\mu_{I,R} > 0$, choose k with $2^{-k} < \mu_{I,R}/2$; then $a_{I,R,k} \geq \mu_{I,R} - 2^{-(k+3)} > 2^{-k}$, so the relation fails. Finally, continuity and attainment on the box give

$$I \in \textsc{Softmax-Exact-Fit} \quad \Longleftrightarrow \quad \exists R \in \mathbb{N}\ \forall k \in \mathbb{N}\ \mathcal{R}(I, R, k). \quad (9)$$

Indeed, every real parameter tuple lies in some integer box, while $\mu_{I,R} = 0$ supplies an actual zero in that box. Equation (9) is a Σ_2^0 definition. It does not imply unconditional decidability or membership in NP.

For completeness, the precise conditional assumption used above is the real form of Schanuel's conjecture: whenever $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$ are linearly independent over $\mathbb{Q}$,

$$\operatorname{trdeg}_{\mathbb{Q}} \mathbb{Q}(\alpha_1, \ldots, \alpha_m, e^{\alpha_1}, \ldots, e^{\alpha_m}) \geq m.$$

Macintyre and Wilkie proved under this assumption that the complete first-order theory of $\mathbb{R}_{\exp}$ is decidable (Macintyre & Wilkie, 1996). Applying that decision procedure to the existential sentence constructed above proves the last assertion of Theorem 2.

The published general bound $\exists\mathbb{R}_{\exp} \subseteq \Sigma_3^0$ is due to Hankala et al. (2024, Theorem 7 and Corollary 8). A broader Σ_2^0 bound is announced in the public abstract of Tanzini (2025). The argument above proves the upper bound for softmax exact fitting directly.

D THE HARDMAX CERTIFICATE

Proof of NP-completeness in Theorem 4. It remains to encode the exact-fit constraints as rational linear (in)equalities.

For each sample $X_t \in \mathbb{Q}^{n\times p_t}$ at $d_{\mathrm{K}} = 1$, the score matrix factorizes as

$$X_t^\top W_{\mathrm{Q}}^\top W_{\mathrm{K}} X_t \;=\; (W_{\mathrm{Q}} X_t)^\top (W_{\mathrm{K}} X_t) \;\in\; \mathbb{R}^{p_t\times p_t},$$

a matrix of rank at most one whose ith column is $(W_{\mathrm{K}}X_t)_i \cdot W_{\mathrm{Q}}X_t$. The hardmax pattern of column i, i.e. the set of coordinates achieving the column maximum (and hence receiving equal hardmax weight $1/k$, with k the size of that set), depends only on the sign of $(W_{\mathrm{K}}X_t)_i$ and on the argmax (resp. argmin) of $W_{\mathrm{Q}}X_t$:

- If $(W_{\mathrm{K}}X_t)_i > 0$: column i of $\mathrm{hardmax}(X_t^\top W_{\mathrm{Q}}^\top W_{\mathrm{K}} X_t)$ assigns weight $1/k$ to each of the k coordinates achieving $\max_j (W_{\mathrm{Q}}X_t)_j$, and 0 elsewhere.
- If $(W_{\mathrm{K}}X_t)_i < 0$: same with $\arg\min$ in place of $\arg\max$ (and the k coordinates achieving $\min_j (W_{\mathrm{Q}}X_t)_j$).
- If $(W_{\mathrm{K}}X_t)_i = 0$: column i is uniform with weight $1/p_t$ in every coordinate.

Thus a certificate only needs to record a finite sign-and-order pattern for each column.

In other words, we can record (i) the sign of $(W_{\mathrm{K}}X_t)_i$, and (ii) the equality/inequality pattern among the coordinates of $W_{\mathrm{Q}}X_t$ that specifies the maximum and minimum. Once those patterns are known, $\mathrm{hardmax}(X_t^\top W_{\mathrm{Q}}^\top W_{\mathrm{K}} X_t)$ is determined. For example, if $(W_{\mathrm{K}}X_t)_i > 0$, and the first and second coordinates of $W_{\mathrm{Q}}X_t$ achieve maximum, then W_{Q} satisfies $(W_{\mathrm{Q}}X_t)_1 = (W_{\mathrm{Q}}X_t)_2, (W_{\mathrm{Q}}X_t)_1 > (W_{\mathrm{Q}}X_t)_3, \ldots, (W_{\mathrm{Q}}X_t)_1 > (W_{\mathrm{Q}}X_t)_{p_t}$. Furthermore, for any $W_{\mathrm{Q}}, W_{\mathrm{K}}$ satisfying the above conditions, the ith column of $\mathrm{hardmax}(X_t^\top W_{\mathrm{Q}}^\top W_{\mathrm{K}} X_t)$ must be $(\frac{1}{2}, \frac{1}{2}, 0, \ldots, 0)$.

Once the pattern is fixed, the matrix $H_t := \mathrm{hardmax}(X_t^\top W_{\mathrm{Q}}^\top W_{\mathrm{K}} X_t)$ becomes a constant rational matrix, and the exact-fit equation $W_{\mathrm{V}} X_t H_t = Y_t$ is rational-linear in W_{V}. Aggregating across all t, we obtain a finite rational system of linear equations and inequalities in Θ, satisfied by the original real solution by construction. By Lemma 5, there exist rational parameters Θ of polynomial bit length that satisfy the system of equations and inequalities, and solve HARDMAX-EXACT-FIT. □

E PROOF OF APPROXIMATE-FIT HARDNESS

We prove Theorem 5 on the scalar-target datasets from the main reductions. Throughout the soundness proof, let $M \in \mathbb{R}^{n\times n}$ be arbitrary and define the attention output by replacing $W_{\mathrm{Q}}^\top W_{\mathrm{K}}$ in (1) and (2) with M. We write $L_\tau(M, W_{\mathrm{V}})$ and $E_\tau(M, W_{\mathrm{V}})$ for the resulting errors, put $s(a, b) = a^\top M b$, and let $v(a) = W_{\mathrm{V}} a \in \mathbb{R}$. Consequently the argument already includes every matrix of the form $W_{\mathrm{Q}}^\top W_{\mathrm{K}}$, regardless of d_{K}.

E.1 SOFTMAX

Lemma 6 (Softmax soundness with error)**.** *For the softmax dataset of Section 3.1, if $E_{\mathrm{S}}(M, W_{\mathrm{V}}) \le \eta$ for some $\eta \le 10^{-3}$, then ϕ is 1-in-3 satisfiable.*

Proof. Retain the odds $\alpha_i, \beta_i, \gamma_i$ from (4), and set

$$\rho(x_i) = \alpha_i, \qquad \rho(\neg x_i) = \beta_i, \qquad \delta_j = e^{s(F,F) - s(D_j, F)}.$$

Linearity of $s(\cdot, F)$ and the definitions of P_i, D_j give the exact identities

$$\gamma_i = \alpha_i \beta_i, \qquad \delta_j = \prod_{r=1}^{3} \rho(\ell_{j,r}). \tag{10}$$

For any sample $[F, U]$ in Groups 3–6, put $\omega_U = e^{s(F,F)-s(U,F)}$. Its first output column is

$$\frac{\omega_U v(F) + v(U)}{\omega_U + 1}. \tag{11}$$

Groups 1 and 2 first give

$$|v(F)| \le \eta, \qquad |v(P_i) - 5| \le \eta. \tag{12}$$

Applying (11) to Groups 3 and 4 and multiplying by the positive denominators yields

$$|v(L_i) - (1 + \alpha_i)| \le \eta(2\alpha_i + 1), \tag{13}$$

$$|v(\bar{L}_i) - (1 + \beta_i)| \le \eta(2\beta_i + 1). \tag{14}$$

Indeed, one copy of $\eta(\alpha_i + 1)$ comes from the output error and the remaining $\alpha_i\eta$ from $v(F)$, and similarly for β_i. Using the exact relation $v(P_i) = v(L_i) + v(\bar{L}_i) - v(F)$, and writing $S_i = \alpha_i + \beta_i$, equations (12)–(14) give

$$|S_i - 3| \le \eta(2S_i + 4). \tag{15}$$

Since $\eta \le 10^{-3} < 1/24$, rearranging the upper side of this inequality gives

$$S_i \le \frac{3 + 4\eta}{1 - 2\eta} \le 4, \qquad |S_i - 3| \le 12\eta.$$

The first column of Group 5 has target $5/3$. Combining its error bound with (12) gives

$$\left|5 - \frac{5}{3}(\gamma_i + 1)\right| \le 2\eta(\gamma_i + 1), \qquad |\gamma_i - 2| \le \frac{6}{5}\eta(\gamma_i + 1).$$

Here the three contributions on the first right-hand side are $|v(P_i) - 5| \le \eta$, at most $\eta(\gamma_i + 1)$ from the output residual after clearing the denominator, and $\gamma_i|v(F)| \le \gamma_i\eta$. By (10) and AM–GM, $\gamma_i = \alpha_i\beta_i \le S_i^2/4 \le 4$, and hence

$$|\gamma_i - 2| \le 6\eta. \tag{16}$$

The numbers α_i, β_i are therefore the roots of $z^2 - S_i z + \gamma_i$. Write $S_i = 3 + s_i$ and $\gamma_i = 2 + g_i$. Its discriminant satisfies

$$\Delta_i = S_i^2 - 4\gamma_i = 1 + 6s_i - 4g_i + s_i^2, \qquad |\Delta_i - 1| \le 96\eta + 144\eta^2 \le 97\eta.$$

Here the last inequality uses $\eta \le 1/144$. Since $\eta \le 10^{-3} < 1/970$, we have $\Delta_i \ge 0.9$, and thus

$$|\sqrt{\Delta_i} - 1| = \frac{|\Delta_i - 1|}{\sqrt{\Delta_i} + 1} \le \frac{97}{1 + \sqrt{0.9}}\eta \le 52\eta.$$

Comparing the roots $(S_i \pm \sqrt{\Delta_i})/2$ with 2 and 1 shows that each distance is at most $(12\eta + 52\eta)/2 = 32\eta$. Thus one of α_i, β_i is within 32η of 2, and the other is within 32η of 1. These neighborhoods are disjoint because $\eta < 1/64$. Declare x_i true precisely when α_i is the root near 2, and let $\widehat{\rho}(\ell) \in \{1, 2\}$ be the corresponding rounded literal odds. Then

$$|\rho(\ell) - \widehat{\rho}(\ell)| \le 32\eta, \qquad \rho(\ell) \le 2 + 32\eta. \tag{17}$$

Fix a clause c_j and let k_j be its number of true literals under this assignment. With $q = 32\eta$, the product identity in (10) and (17) imply

$$\begin{aligned} |\delta_j - 2^{k_j}| &\le 3q\,2^2 + 3q^2\,2 + q^3 \\ &= 384\eta + 6144\eta^2 + 32768\eta^3 \le 400\eta. \end{aligned} \tag{18}$$

For the last inequality, after division by $\eta > 0$, the higher-order part is at most $6144(10^{-3}) + 32768(10^{-3})^2 < 16$; the case $\eta = 0$ is immediate. Likewise, (13), (14), and (17) give, for every literal ℓ,

$$|v(T(\ell)) - (1 + \widehat{\rho}(\ell))| \le 32\eta + \eta(5 + 64\eta).$$

Since $D_j = \sum_r T(\ell_{j,r}) - 2F$, it follows that

$$|v(D_j) - (6 + k_j)| \le 3(32\eta) + 3\eta(5 + 64\eta) + 2\eta = 113\eta + 192\eta^2 \le 114\eta. \tag{19}$$

Finally, the first output column in Group 6 has target $7/3$, so

$$\left|\delta_j v(F) + v(D_j) - \frac{7}{3}(\delta_j + 1)\right| \le \eta(\delta_j + 1). \tag{20}$$

Write $\delta_j = 2^{k_j} + r_j$ and $v(D_j) = 6 + k_j + r'_j$. Equations (18) and (19) give $|r_j| \le 400\eta$ and $|r'_j| \le 114\eta$. Moreover, $\delta_j \le 8.4$, $2^{k_j} \le 8$, and $|v(F) - 7/3| < 5/2$. Isolating the ideal clause expression in (20) uses the exact decomposition

$$\delta_j v(F) + v(D_j) - \frac{7}{3}(\delta_j + 1) = \left((6 + k_j) - \frac{7}{3}(2^{k_j} + 1)\right) + 2^{k_j} v(F) + r_j \left(v(F) - \frac{7}{3}\right) + r'_j.$$

It therefore gives

$$\begin{aligned}\left|(6 + k_j) - \frac{7}{3}(2^{k_j} + 1)\right| &\le 10\eta + 8\eta + (400\eta)\frac{5}{2} + 114\eta \\ &= 1132\eta.\end{aligned}$$

Equivalently,

$$|3(6 + k_j) - 7(2^{k_j} + 1)| \le 3396\eta \le 3.396.$$

For $k = 0, 1, 2, 3$, the expression inside the absolute value is respectively $4, 0, -11, -36$. Only $k_j = 1$ is possible. Thus every clause has exactly one true literal, proving the lemma. □

It follows immediately that an unsatisfiable instance has $E_{\mathrm{S}} > 10^{-3}$ for every parameter choice. Also, $E_{\mathrm{S}} \le \sqrt{L_{\mathrm{S}}}$, because every squared coordinate residual is a nonnegative summand of the total loss. Hence $L_{\mathrm{S}} \le 10^{-6}$ would satisfy the lemma's hypothesis, so every parameter choice on an unsatisfiable instance has $L_{\mathrm{S}} > 10^{-6}$ and $\mathrm{OPT}_{\mathrm{S}} \ge 10^{-6}$.

E.2 HARDMAX

Lemma 7 (Hardmax soundness with error)**.** *For the hardmax dataset of Section 4.1, either of the conditions*

$$L_{\mathrm{H}}(M, W_{\mathrm{V}}) < \frac{1}{5}, \qquad E_{\mathrm{H}}(M, W_{\mathrm{V}}) < \frac{1}{4}$$

implies that ϕ is 1-in-3 satisfiable.

Proof. Write $a_i = v(L_i)$ and $b_i = v(\bar{L}_i)$. We first use the total-loss hypothesis directly. Let

$$\varepsilon_i^P = |a_i + b_i - 1|,$$

let $\varepsilon_{i,1}, \varepsilon_{i,2}$ be the two absolute residuals of $[L_i, \bar{L}_i] \mapsto (1, 0)$, and let $\varepsilon_j^D = |v(D_j) - 1|$. Their squares sum to L_{H}.

Suppose a column of $[L_i, \bar{L}_i]$ is tied. Its output is $(a_i + b_i)/2$. Put $e = \varepsilon_i^P$. If $e \le 1$, the error of the tied column against either target 0 or 1 is at least $(1 - e)/2$. The singleton coordinate and this distinct column therefore contribute at least

$$e^2 + \left(\frac{1 - e}{2}\right)^2 = \frac{1}{5} + \frac{5}{4}\left(e - \frac{1}{5}\right)^2 \ge \frac{1}{5}.$$

If $e \ge 1$, the singleton coordinate alone contributes at least 1. Thus $L_{\mathrm{H}} < 1/5$ rules out every tie. If the two columns instead had the same strict winner with value w, their two residuals would contribute

$$(w - 1)^2 + w^2 = \frac{1}{2} + 2\left(w - \frac{1}{2}\right)^2 \ge \frac{1}{2},$$

also impossible. Hence the two columns have distinct strict winners.

Declare x_i true exactly when the first column selects L_i. The first-column residual measures the selected true literal's distance from 1, and the second-column residual measures the selected false literal's distance from 0. If $\widehat{v}(T(\ell)) \in \{0, 1\}$ is the induced Boolean value, then

$$|v(T(\ell)) - \widehat{v}(T(\ell))| = \varepsilon_{i(\ell), c(\ell)}, \qquad c(\ell) = \begin{cases} 1, & \ell \text{ is true,} \\ 2, & \ell \text{ is false.} \end{cases} \tag{21}$$

For the three pairwise distinct literals of a clause, the three pairs $(i(\ell), c(\ell))$ are distinct: opposite literals of the same variable have opposite truth values and hence use different columns.

Let $k_j \in \{0, 1, 2, 3\}$ be the number of true literals in c_j. Since $v(D_j)$ is exactly the sum of its three literal values,

$$|k_j - 1| \le \sum_{r=1}^{3} \varepsilon_{i(\ell_{j,r}),c(\ell_{j,r})} + \varepsilon_j^D.$$

These are four distinct residual coordinates. Cauchy–Schwarz therefore gives

$$|k_j - 1| \le 2\sqrt{L_{\text{H}}} < \frac{2}{\sqrt{5}} < 1.$$

As k_j is an integer, $k_j = 1$.

For the coordinatewise statement, now put $\eta = E_{\text{H}} < 1/4$. Group 1 gives $|a_i + b_i - 1| \le \eta$. A tied column would output $(a_i + b_i)/2$, whose error against either target is at least $1/2 - \eta/2 > \eta$, impossible. If both columns had the same strict winner w, then $|w - 1| \le \eta$ and $|w| \le \eta$, implying $1 \le 2\eta$, again impossible. The two winners are therefore distinct, and the same assignment as above puts every literal value within η of its Boolean value. The clause coordinate is also within η of 1, so

$$|k_j - 1| \le 4\eta < 1.$$

Again $k_j = 1$ in every clause. This proves both assertions. □

Contrapositively, every parameter choice for an unsatisfiable hardmax instance has $L_{\text{H}} \ge 1/5$ and $E_{\text{H}} \ge 1/4$, so $\text{OPT}_{\text{H}} \ge 1/5$.

E.3 Parameter dimensions and completion of the proof

The two lemmas were proved for an arbitrary $M \in \mathbb{R}^{n\times n}$. Given parameters of any key dimension $d_{\text{K}} \ge 1$, substituting $M = W_{\text{Q}}^{\top} W_{\text{K}}$ therefore proves the same soundness statement without any rank assumption. In the yes-case, the rank-one parameters from the exact completeness proofs remain valid for every d_{K} after adding zero rows to W_{Q} and W_{K}. The padding is used only for this completeness direction.

For $d_{\text{V}} > 1$, replicate every scalar target in all d_{V} rows, as in Section 4.2. If $W_{\text{V}}^{(r)}$ denotes row r of W_{V}, then the attention weights do not depend on W_{V}, and hence

$$\begin{aligned} L_\tau(\Theta; \overline{\mathcal{D}}) &= \sum_{r=1}^{d_{\text{V}}} L_\tau((W_{\text{Q}}, W_{\text{K}}, W_{\text{V}}^{(r)}); \mathcal{D}), \\ E_\tau(\Theta; \overline{\mathcal{D}}) &= \max_r E_\tau((W_{\text{Q}}, W_{\text{K}}, W_{\text{V}}^{(r)}); \mathcal{D}). \end{aligned} \tag{22}$$

Thus any total-loss or coordinatewise upper bound for the full output holds for its first row, where the scalar soundness lemmas apply. In the yes-case, replicate the scalar W_{V} row to preserve the exact fit. This proves Theorem 5 for every fixed positive $d_{\text{K}}, d_{\text{V}}$.

E.4 Two tokens and bounded parameter entries

The exact and approximate hardness results remain valid when every sample has exactly two tokens. Replace each singleton sample $([x], y)$ by $([x, x], [y, y])$, keeping every two-token sample unchanged. For either attention variant and every parameter tuple, both output columns of $[x, x]$ equal $W_{\text{V}}x$. Thus the transformation preserves exact fitting and maximum coordinate error. It doubles each singleton's contribution to total squared loss and leaves the other contributions unchanged, so all the lower bounds in Theorem 5 remain valid. The number of samples and the input and target alphabets are unchanged.

The same hardness results hold if every entry of $W_{\text{Q}}, W_{\text{K}}, W_{\text{V}}$ is required to lie in $[-3, 3]$ for softmax, or in $[-1, 1]$ for hardmax. The displayed completeness witnesses already satisfy these bounds; adding zero query and key rows and replicating the value row preserves them. Restricting the parameter space cannot weaken soundness, which holds for arbitrary real parameters. These restrictions

can therefore be imposed together with exactly two tokens per sample, while retaining the original targets, bounded in magnitude by 5 for softmax and 1 for hardmax. The bounds are entrywise; they do not give a dimension-independent bound on the Frobenius norms of the parameter matrices.

E.5 Target scaling and mean squared loss

For $r > 0$, let $\mathcal{D}^{(r)}$ be obtained from $\mathcal{D}$ by replacing each Y_t with rY_t. Since attention weights do not depend on W_{V},

$$A_\tau(X; W_{\mathrm{Q}}, W_{\mathrm{K}}, rW_{\mathrm{V}}) = rA_\tau(X; W_{\mathrm{Q}}, W_{\mathrm{K}}, W_{\mathrm{V}}).$$

Scaling W_{V} is a bijection on the unrestricted parameter space, so $\mathrm{OPT}_\tau(\mathcal{D}^{(r)}) = r^2\,\mathrm{OPT}_\tau(\mathcal{D})$, and an attained exact fit remains attained. For $N \geq 1$, choosing the integer $r = N$ gives

$$\frac{\mathrm{OPT}_\tau(\mathcal{D}^{(N)})}{N} = N\,\mathrm{OPT}_\tau(\mathcal{D}).$$

Thus either constant total-loss gap in Theorem 5 yields a constant sample-mean gap after this polynomial-time transformation. Each target entry needs only $O(\log N)$ additional bits, and the input sequences are unchanged. This observation allows targets and value parameters to grow with N; it does not preserve the uniform bounds in Appendix E.4.

F Proof of mean-squared-loss hardness

We prove Theorem 6 by repeating samples from the original reductions. The repetitions ensure that every violated clause contributes a fixed amount to the total loss, while the number of samples remains proportional to the number of clauses.

A gap version of 1-in-3-SAT. By Håstad (2001, Theorem 6.5), it is NP-hard to distinguish satisfiable 3-CNF formulas from formulas for which every assignment violates at least $1/16$ of the clauses. This is the theorem's perfect-completeness case with $\epsilon = 1/16$. To obtain a 1-in-3-SAT instance, replace each disjunction $(\ell_1 \vee \ell_2 \vee \ell_3)$ by four exactly-one constraints

$$(u_1, u_2, u_3), \qquad (u_r, \neg\ell_r, z_r), \quad r = 1, 2, 3,$$

where the six auxiliary variables are fresh for each disjunction. The last three constraints say $z_r = \ell_r - u_r$ on Boolean truth values, so they allow precisely $u_r \leq \ell_r$. Together with $u_1 + u_2 + u_3 = 1$, they are satisfiable exactly when the disjunction is true. If the disjunction is false, at least one of the four constraints is violated, whatever the auxiliary values. Thus the resulting formula ϕ has m clauses, each on three distinct variables, and either is satisfiable or every assignment violates at least $m/64$ clauses. Remove unused variables; then the number of variables is at most $3m$. We use nonempty source instances, so $m \geq 1$.

The repeated samples. For softmax, use the vectors and targets of Section 3.1. For each clause c_j, include one copy of $[F] \mapsto 0$ and its clause sample $[F, D_j] \mapsto (7/3, 14/3)$. For each of its three variables, include one copy of each of the four variable samples in Groups 2–5. This gives $2 + 3 \cdot 4 = 14$ samples per clause and $N = 14m$ in total. For hardmax, use Section 4.1: each clause receives the two samples in Groups 1–2 for each of its three variables and its own sample $[D_j] \mapsto 1$. Here $N = 7m$. The dataset is an indexed list; repeated samples are retained as separate observations. The sample groups belonging to different clauses therefore occupy disjoint positions in the list, even when their inputs and targets coincide. All entries retain their original constant bit length, and the construction has polynomial size. A satisfying assignment gives an exact fit using the same scalar parameters as in the original completeness proofs.

Softmax soundness. Fix any real score matrix M and scalar value row, and define one assignment by declaring x_i true exactly when $\alpha_i \geq 3/2$, with the odds from (4). Consider a clause whose 14 samples have total squared loss at most 10^{-6}. Every coordinate residual in these samples is then at most $\eta = 10^{-3}$. The variable estimates in the proof of Lemma 6 use only the sample $[F]$ and that variable's four samples. They give

$$|\alpha_i + \beta_i - 3| \leq 12\eta, \qquad |\alpha_i\beta_i - 2| \leq 6\eta,$$

and place the two odds within 32η of 1 and 2, one near each. These neighborhoods lie on opposite sides of $3/2$, so their rounding agrees with the single assignment just defined. For this clause, the remaining estimates in the same proof use only these variable samples, $[F]$, and the clause sample. If k_j is its number of true literals, they yield

$$|3(6 + k_j) - 7(2^{k_j} + 1)| \leq 3396\eta < 4,$$

which forces $k_j = 1$. Consequently every clause violated by the assignment has loss greater than 10^{-6} in its own 14 samples. In the no-case at least $m/64$ clauses are violated. Summing over their disjoint sample positions gives

$$\frac{L_{\mathrm{S}}}{N} \geq \frac{10^{-6}}{64 \cdot 14} = \frac{1}{896\,000\,000}.$$

The bound holds for every real M and value row, and hence also for the infimum of the loss.

Hardmax soundness. Fix any real M and scalar value row. Define one assignment by declaring x_i true when the first column of $[L_i, \bar{L}_i]$ has strict winner L_i, and false otherwise. Consider a clause whose seven samples have total squared loss less than $1/5$. The proof of Lemma 7 excludes a tied variable column using only that column and its singleton sample; it excludes identical strict winners using only the two variable columns. These samples are present for all three variables of this clause. Thus both columns for every such variable have distinct strict winners, and their Boolean values agree with the assignment just defined. The three literal residuals and the clause residual occupy four distinct coordinates. Writing $\mathcal{L}_j$ for the total loss of these seven samples, the same rounding argument gives

$$|k_j - 1| \leq 2\sqrt{\mathcal{L}_j} < 2/\sqrt{5} < 1.$$

Hence $k_j = 1$. Every violated clause therefore contributes at least $1/5$ in its seven samples. The no-case gives

$$\frac{L_{\mathrm{H}}}{N} \geq \frac{1}{64 \cdot 5 \cdot 7} = \frac{1}{2240}.$$

Again this holds for arbitrary real parameters and their infimum.

The remaining restrictions. For larger fixed key dimension, pad the scalar completeness parameters with zero rows; soundness already holds for arbitrary M. For larger value dimension, replicate the targets and the completeness value row. Applying scalar soundness to one output row preserves the same mean-loss lower bound. Thus the smaller of the two displayed constants may be used as c in Theorem 6 for all fixed positive dimensions. The completeness parameters satisfy the entrywise bounds in Appendix E.4. Replacing each singleton by two identical tokens, with its target repeated, preserves N, exact fitting, and these bounds, and cannot decrease total loss. This gives exactly two tokens per sample with the same mean-loss constants. Finally, if targets in $[0, 1]$ are desired, divide the softmax targets and the completeness value row by 5; the correspondence in Appendix E.5 divides the mean-loss gap by 25. The hardmax targets already lie in $\{0, 1\}$.

G CAUSAL ATTENTION WITH ONE SCALAR TARGET

For two tokens $X = [a, b]$, a causal mask allows the first output to attend only to a, and the second to attend to both a and b. We supervise only the second output, which is unchanged if the mask is omitted. All three rows $W_{\mathrm{Q}}, W_{\mathrm{K}}, W_{\mathrm{V}}$ are trainable. As before, write $v(x) = W_{\mathrm{V}}x$ and $s(a, b) = (W_{\mathrm{Q}}a)(W_{\mathrm{K}}b)$. The supervised output is

$$f_\Theta([a, b]) = \frac{v(a)e^{s(a,b)} + v(b)e^{s(b,b)}}{e^{s(a,b)} + e^{s(b,b)}}. \tag{23}$$

Given a nonempty indexed dataset $(X_t, y_t)_{t=1}^N$ with $X_t \in \mathbb{Q}^{n\times 2}$ and $y_t \in \mathbb{Q}$, exact fitting asks whether real rows $W_{\mathrm{Q}}, W_{\mathrm{K}}, W_{\mathrm{V}}$ exist with $f_\Theta(X_t) = y_t$ for every t. The mean squared loss is $N^{-1}\sum_t (f_\Theta(X_t) - y_t)^2$. Also, $f_\Theta([x, x]) = v(x)$ for every x. With $F = e_1$, write $q(x) = s(x, F) = (W_{\mathrm{Q}}x)(W_{\mathrm{K}}F)$, which is linear in x, and put $\omega_U = e^{q(F)-q(U)}$. Then

$$f_\Theta([U, F]) = \frac{v(U) + \omega_U v(F)}{1 + \omega_U}. \tag{24}$$

Proposition 2 (One scalar target per sequence)**.** *For the causal model* (23) *with* $d_{\mathrm{K}} = d_{\mathrm{V}} = 1$*, exact fitting is NP-hard. It is also NP-hard to distinguish datasets admitting an exact fit from those whose infimum mean squared loss is at least* $1/(896\,000\,000)$*. Every sample has exactly two tokens and one scalar target, with input entries in* $\{-2, -1, 0, 1\}$ *and targets in* $\{0, 1, 5/3, 7/3, 5\}$*. The yes-instances admit fitting parameters with entries in* $[-3, 3]$*.*

Construction. Use the gap version of 1-in-3-SAT established in Appendix F: a formula ϕ has $m \geq 1$ clauses, each on three distinct variables, and either is satisfiable or every assignment violates at least $m/64$ clauses. Let d be its number of variables and $n = 2d + 1$. Use the basis vectors $F, L_i, \bar{L}_i$ and the vectors $P_i = L_i + \bar{L}_i - F$ and $D_j = \sum_{r=1}^{3} T(\ell_{j,r}) - 2F$ from the softmax reduction. For each clause c_j, include $[F, F] \mapsto 0$ and $[D_j, F] \mapsto 7/3$, together with the four samples

$$[P_i, P_i] \mapsto 5, \quad [P_i, F] \mapsto \frac{5}{3},$$
$$[L_i, F] \mapsto 1, \quad [\bar{L}_i, F] \mapsto 1$$

for each of its three variables. This gives 14 samples per clause. Every label specifies only the last scalar output. Samples belonging to different clauses occupy distinct positions in the indexed dataset, even if their inputs coincide. Hence $N = 14m$. The input entries belong to $\{-2, -1, 0, 1\}$ and the targets to $\{0, 1, 5/3, 7/3, 5\}$. The construction has polynomial size.

Completeness. For a satisfying assignment, choose $W_{\mathrm{K}} = e_1^\top$, so $q(x) = W_{\mathrm{Q}} x$, and set $q(F) = v(F) = 0$. For each literal vector, set $(q(T(\ell)), v(T(\ell))) = (-\log 2, 3)$ if ℓ is true, and $(0, 2)$ otherwise. Since these vectors together with F form a basis, this specifies $W_{\mathrm{Q}}, W_{\mathrm{V}}$. All three rows have entries in $[-3, 3]$. Exactly one of a variable and its negation is true, so $v(P_i) = 5$ and $\omega_{P_i} = 2$. Equation (24) gives output 1 on each literal sample and $5/3$ on $[P_i, F]$. Each clause has exactly one true literal, so $v(D_j) = 7$ and $\omega_{D_j} = 2$, giving $7/3$. The duplicate-token samples give $v(F) = 0$ and $v(P_i) = 5$. All labels are fitted exactly.

Local soundness. Fix arbitrary real rows $W_{\mathrm{Q}}, W_{\mathrm{K}}, W_{\mathrm{V}}$, and define

$$\alpha_i = e^{q(F) - q(L_i)}, \qquad \beta_i = e^{q(F) - q(\bar{L}_i)}, \qquad \rho(x_i) = \alpha_i, \quad \rho(\neg x_i) = \beta_i.$$

Linearity of q gives

$$\omega_{P_i} = \alpha_i \beta_i, \qquad \omega_{D_j} = \prod_{r=1}^{3} \rho(\ell_{j,r}).$$

Declare x_i true precisely when $\alpha_i \geq 3/2$. This is one assignment determined by the shared parameters for the entire dataset. Suppose the 14 samples belonging to a clause c_j have total squared loss at most 10^{-6}. Every residual in this group has absolute value at most $\eta = 10^{-3}$.

For each of its three variables, the duplicate-token and literal samples, together with (24), imply

$$|v(F)| \leq \eta, \qquad |v(P_i) - 5| \leq \eta,$$
$$|v(L_i) - (1 + \alpha_i)| \leq \eta(2\alpha_i + 1), \qquad |v(\bar{L}_i) - (1 + \beta_i)| \leq \eta(2\beta_i + 1).$$

Writing $S_i = \alpha_i + \beta_i$ and using $v(P_i) = v(L_i) + v(\bar{L}_i) - v(F)$ gives

$$|S_i - 3| \leq \eta(2S_i + 4), \qquad S_i \leq \frac{3 + 4\eta}{1 - 2\eta} \leq 4, \qquad |S_i - 3| \leq 12\eta.$$

The sample $[P_i, F] \mapsto 5/3$, with $\gamma_i = \alpha_i \beta_i$, gives

$$\left| 5 - \frac{5}{3}(\gamma_i + 1) \right| \leq 2\eta(\gamma_i + 1), \qquad |\gamma_i - 2| \leq \frac{6}{5}\eta(\gamma_i + 1) \leq 6\eta,$$

where $\gamma_i \leq S_i^2/4 \leq 4$. The positive numbers α_i, β_i are the roots of $z^2 - S_i z + \gamma_i$. Its discriminant Δ_i satisfies

$$|\Delta_i - 1| \leq 96\eta + 144\eta^2 \leq 97\eta, \qquad \Delta_i \geq 0.9, \qquad |\sqrt{\Delta_i} - 1| \leq 52\eta.$$

The quadratic formula therefore places one root within 32η of 1 and the other within 32η of 2. These intervals are on opposite sides of $3/2$, so this rounding agrees with the assignment already defined. If $\widehat{\rho}(\ell)$ is 2 for a true literal and 1 for a false literal, then $|\rho(\ell) - \widehat{\rho}(\ell)| \le 32\eta$ for all literals in c_j.

Let k be the number of true literals in this clause, and write $\delta = \omega_{D_j}$. Expanding the product of its three odds gives

$$|\delta - 2^k| \le 384\eta + 6144\eta^2 + 32768\eta^3 \le 400\eta.$$

Each literal value is within $32\eta + \eta(5 + 64\eta)$ of $1 + \widehat{\rho}(\ell)$. Linearity of v therefore gives

$$|v(D_j) - (6 + k)| \le 113\eta + 192\eta^2 \le 114\eta.$$

Finally, the clause sample implies

$$\left|\delta v(F) + v(D_j) - \frac{7}{3}(1 + \delta)\right| \le \eta(1 + \delta).$$

Put $r = \delta - 2^k$ and $r' = v(D_j) - (6 + k)$, and use the identity

$$\delta v(F) + v(D_j) - \frac{7}{3}(1 + \delta) = 6 + k - \frac{7}{3}(1 + 2^k) + 2^k v(F) + r\left(v(F) - \frac{7}{3}\right) + r'.$$

Since $\delta \le 8.4$, $2^k \le 8$, and $|v(F) - 7/3| < 5/2$, it follows that

$$\left|6 + k - \frac{7}{3}(1 + 2^k)\right| \le 10\eta + 8\eta + 1000\eta + 114\eta = 1132\eta.$$

Thus $|3(6+k) - 7(1+2^k)| \le 3396\eta < 4$. The expression inside the absolute value is $4, 0, -11, -36$ for $k = 0, 1, 2, 3$, respectively. Consequently $k = 1$.

Mean squared loss. Every clause violated by the single assignment defined above therefore contributes more than 10^{-6} to the loss in its own 14 samples. In the no-case at least $m/64$ clauses are violated. The sample groups are disjoint, so every choice of the real trainable rows satisfies

$$\frac{1}{N}\sum_{t=1}^{N}\big(f_\Theta(X_t) - y_t\big)^2 > \frac{10^{-6}}{64 \cdot 14} = \frac{1}{896\,000\,000}.$$

Taking the infimum preserves the weak lower bound in the proposition. The same reduction proves NP-hardness of exact fitting, since every yes-instance has an attained exact fit and no no-instance does.