hexagon:2611.00132v1

Fast Attention after SETH: The Exact Exponent and the Hypotheses Behind It

Xiaoyu Li*

Abstract

Alman and Song showed that attention with head dimension $O(\log n)$ can be approximated in almost linear time when the entries are $o(\sqrt{\log n})$, and needs quadratic time at $\Theta(\sqrt{\log n})$ under the Strong Exponential Time Hypothesis. In October 2026 the Exponential Time Hypothesis was refuted, and with it SETH. The quadratic lower bound now rests on the hardness of Orthogonal Vectors in dimension $c \log n$ (LDOV), which no open hypothesis about satisfiability is known to imply. We measure an instance by its spread W, the largest gap between two logits of a row. At every nontrivial constant precision and in every dimension between $\omega(\log n)$ and $n^{o(1)}$, we show that under LDOV the time exponent of attention is exactly $1+\min\{W/\ln n, 1\}$: uniform sampling of keys is optimal, the cost is $\min\{n\, e^W, n^2\}$ up to factors $n^{o(1)}$, and the transition to quadratic time sits at $W = \ln n$, the spread at which one key can outweigh all others. At polynomial precision and in dimension $O(\log n)$ the threshold becomes $\Theta(\log n)$, with a constant that tends to zero as the precision grows, and the dimension moves it: at precision $n^{-1-O(1/\log\log n)}$, spread $O(\log n/\log\log n)$ is hard in dimension $(\log n)^{1+\Omega(1)}$ but easy in dimension $O(\log n)$. From satisfiability we obtain hardness at spread $(1+\theta)\ln n$ in dimension $n^{o(1)}$, provided the hard CNF formulas have N^{c_0} clauses for some $c_0 < 2$, and we show that the construction behind this bound, which builds one instance for each proof of a Merlin–Arthur protocol for disjointness, needs $c_0 < 2$ for every such protocol. For the entry bound, every reduction that embeds Orthogonal Vectors vector by vector, with polynomial blowup and logit gap $\Omega(\log n)$, yields a Merlin–Arthur protocol for disjointness; by Klauck's lower bound, such reductions cannot derive $B = \Theta(\sqrt{\log n})$ from Orthogonal Vectors in dimension $(\log n)^{1+\Omega(1)}$. We also map the conditional lower bounds for attention in the literature to the hypotheses that now support them, and find two families whose hypotheses are false, one of them in part of its parameter range.

*University of New South Wales, xiaoyulics@gmail.com

Contents

1 Introduction **3**
- 1.1 Results . . . 3
- 1.2 What is new . . . 5

2 Preliminaries **7**
- 2.1 Attention and its parameters . . . 7
- 2.2 Hypotheses after the refutation . . . 7
- 2.3 Imported theorems . . . 8

3 The concentration threshold **9**
- 3.1 Sampling . . . 9
- 3.2 Quadratic time above the threshold . . . 10
- 3.3 The exponent below the threshold . . . 11

4 Precision and dimension move the threshold **12**
- 4.1 Counting modulo $r + 1$ . . . 12
- 4.2 The polynomial method . . . 14
- 4.3 Dimension matters at high precision . . . 14

5 What satisfiability still supports **15**
- 5.1 CNF-SETH in polylogarithmic dimension . . . 15
- 5.2 Clause exponent below two . . . 15
- 5.3 A barrier for the bounded-entry threshold . . . 16

6 Which attention lower bounds survive **17**

7 Related work and open problems **21**

A Roadmap of the appendix **27**

B Proofs for Sections 2 and 3 **28**
- B.1 Sampling . . . 29
- B.2 Quadratic time above the threshold . . . 30
- B.3 The exponent below the threshold . . . 32

C Proofs for Section 4 **34**
- C.1 Counting modulo $r + 1$ . . . 34
- C.2 The polynomial method . . . 37
- C.3 Dimension matters at high precision . . . 38

D Proofs for Section 5 **39**
- D.1 CNF-SETH in polylogarithmic dimension . . . 39
- D.2 Clause exponent below two . . . 40
- D.3 A barrier for the bounded-entry threshold . . . 41

1 Introduction

Attention computes, for each of n queries, an average of n value vectors weighted by the softmax of the query's inner products with n keys. The standard algorithm takes time $\Theta(n^2 d)$. Alman and Song [AS23] found that whether one can do better depends on the size of the entries. With head dimension $d = O(\log n)$ and entries bounded by B, attention can be approximated to precision $1/\mathrm{poly}(n)$ in time $n^{1+o(1)}$ if $B = o(\sqrt{\log n})$, and it needs time $n^{2-o(1)}$ if $B = \Theta(\sqrt{\log n})$, assuming the Strong Exponential Time Hypothesis (SETH). The threshold has been extended to the gradient used in training [AS24a], to tensor attention [AS24b], to large entries in head dimension $2^{\Theta(\log^* n)}$ [GHS+26], to every nontrivial constant precision [HAPM+26], and to multi-head and multi-layer transformers [SXYY26], under SETH or hypotheses derived from it, and it underlies lower bounds for many transformer variants (Section 6).

In October 2026 two works refuted the Exponential Time Hypothesis [Ope26, Cla26]. For every constant k, both count the satisfying assignments of a k-CNF on N variables in time $2^{N/(\log N)^{1/2-o(1)}}$. SETH implies ETH, so SETH is false, and the attention bounds lose the hypothesis they were proved from. The lower bounds themselves are not contradicted: the new algorithms say nothing about Orthogonal Vectors (OV), or its k-party version, the problems these reductions start from. What the refutation removes is the derivation of the bounds from satisfiability. SETH implied the hardness of OV in dimension $c \log n$ through the sparsification lemma, and this route now starts from a false statement, since CNF-SAT with linearly many clauses is subexponential [Li26, Proposition 6.1]. The hardness of OV in dimension $c \log n$, the low-dimensional OV hypothesis (LDOV), is now a hypothesis of its own. The form of SETH that survives, CNF-SETH, gives through Williams' reduction the hardness of OV in dimension $(\log n)^{O(1)}$, and no reduction to dimension $O(\log n)$ is known. The ledger of [Li26, Section 6] records that the attention thresholds now rest on LDOV. Within the same month, the 3SUM and APSP hypotheses were refuted as well [AV26].

This paper determines, under LDOV, the time exponent of attention at constant precision at every spread and in every dimension between $\omega(\log n)$ and $n^{o(1)}$, and the hypotheses on which each part of the answer rests. The answer is sharper than the original threshold, and it separates two readings of it that used to be the same.

1.1 Results

We measure an attention instance by its *spread* W: the largest difference between two logits in the same row. The spread is the normalization-free form of the entry bound: in the normalization of Alman and Song, $W \le 2B^2$. It also has a direct meaning. A row with spread W gives every key at most e^W/n of its weight, so a single key can carry constant weight only when $W \ge \ln n - O(1)$. Precision is additive, and the values lie in $[-1, 1]$.

Theorem A (The exponent at constant precision; Corollaries 3.2 and 3.7, Theorems 3.3 and 3.6, and Remark 3.5)**.** *Fix a precision $\varepsilon \in (0,1)$ and $\kappa > 0$.*

(i) In every dimension $d \le n^{o(1)}$, time $n^{1+\min\{\kappa,1\}+o(1)}$ suffices for attention with spread $\kappa \ln n$: uniform sampling of keys if $\kappa < 1$, direct computation otherwise.

(ii) Under LDOV, for every $\gamma > 0$ there is C such that in dimension $C \log n$, attention with spread $\kappa \ln n$ needs time $n^{1+\min\{\kappa,1\}-\gamma}$, already for values in $\{-1,1\}$, and for values in $\{0,1\}$ if $\varepsilon < 1/2$.

So under LDOV, at constant precision and in every dimension between $\omega(\log n)$ and $n^{o(1)}$, the cost of attention is $n \cdot e^W$ up to factors $n^{o(1)}$ until it reaches n^2, and uniform sampling is optimal. The transition to quadratic time sits at $W = \ln n$, the spread at which one key can first

outweigh all others: for every $\theta > 0$, spread $(1-\theta)\ln n$ takes time $n^{2-\theta+o(1)}$, while under LDOV spread $(1+\theta)\ln n$ has no $O(n^{2-\gamma})$ algorithm for any $\gamma > 0$, already in dimension $C_{\gamma,\theta}\log n$ and at every nontrivial constant precision (Theorem 3.3 and Remark 3.5); in dimension $(\log n)^{O(1)}$, spread $\ln n + O(1)$ suffices (Corollary 3.4). Alman and Song proved hardness at precision $1/\mathrm{poly}(n)$, and Haverbeck, Amo Alonso, Posada-Moreno, Trimpe and Pavone [HAPM+26] at every nontrivial constant precision, both with an unspecified constant C' in $B = C'\sqrt{\log n}$.

Theorem B (Precision and dimension move the threshold; Corollaries 4.2 and 4.5)**.** *Assume LDOV, and let $\gamma \in (0,1)$.*

(i) In dimension $O_\gamma(\log n)$ and at precision $n^{-\beta}$, spread $\kappa \ln n$ can be handled in time $O(n^{2-\gamma})$ for some $\kappa = \kappa_\gamma(\beta) > 0$, while for every $\kappa > 0$ it needs more once $\beta \geq O_\gamma(\log(2+1/\kappa))$. Spread 1, and even entries bounded by 2 in the normalization of [AS23] (Remark 4.3), need time $n^{2-\gamma}$ at precision $n^{-O_\gamma(\log\log n)}$.

(ii) For every $C' > 0$, at precision $n^{-1-O_{\gamma,C'}(1/\log\log n)}$, spread $O_{\gamma,C'}(\log n/\log\log n)$ needs time $n^{2-\gamma}$ in dimension $(\log n)^{1+C'}$, and for every $\delta \in (0, 1/2]$, at precision $\Theta_{\gamma,\delta}(1/n)$, spread $O_{\gamma,\delta}(1)$ needs it in dimension $O_{\gamma,\delta}(n^\delta)$.

So the constant $\ln n$ belongs to constant precision. At polynomial precision the threshold is $\Theta(\log n)$ in dimension $O(\log n)$, with a constant that tends to zero as the precision grows, and, unlike at constant precision, the dimension moves it further: in dimension $O(\log n)$ every spread $o(\log n)$ is easy at every polynomial precision (Proposition 4.4), while under LDOV, in every dimension between $(\log n)^{1+\Omega(1)}$ and $n^{o(1)}$ and at every precision $n^{-\beta}$ with $\beta > 1$, spread $O(\log n/\log\log n)$ is hard (Corollary 4.2(b) and padding). The lower bounds count orthogonal partners modulo $r+1$ through the integer $r^K \sum_b (1+1/r)^{F(a,b)}$, where $F(a,b)$ counts the blocks of coordinates in which a and b intersect. This generalizes the parity argument of Gupta et al. [GHS+26], and the blocks trade dimension for precision.

The next two results concern satisfiability. Both depend on the *clause exponent*: CNF-SETH with exponent c_0 asserts that CNF formulas with N variables and N^{c_0} clauses need time $2^{(1-\gamma)N}$ for every $\gamma > 0$. Exponent 1 is refuted [Li26], and every exponent $c_0 > 1$ is open.

Theorem C (What satisfiability supports; Proposition 5.1, Theorem 5.3, and Remark 3.5)**.**

(i) Under CNF-SETH, for every $\gamma > 0$, attention at precision $1/6$ needs time $n^{2-\gamma}$ in dimension $d = (\log n)^{C_\gamma}$ with spread $d\ln(16n)$. This combines the exact embedding of [GHS+26, Lemma 5.6] with Williams' reduction.

(ii) Under CNF-SETH with clause exponent $c_0 < 2$, for all $\gamma, \theta > 0$ and every nontrivial constant precision, attention with spread $(1+\theta)\ln n$ needs time $n^{2-\gamma}$ in dimension $n^{o(1)}$.

Theorem D (A barrier for the entry bound; Theorem 5.6 and Corollary 5.7)**.** *Let a reduction map each vector of an OV instance in dimension D, for each of 2^p values of an index, to a query or key vector, and separate orthogonal from non-orthogonal pairs by a gap Δ in the logits. Let σ be the product of the largest Euclidean norms of the query vectors and of the key vectors; in the normalization of [AS23] with entries bounded by B, $\sigma \leq B^2$. Then $(\sigma/\Delta)^2 \log(2+\sigma/\Delta) \geq \Omega(D/(p+1))$. In particular, for $D = (\log n)^{c_0}$ with $c_0 > 1$, polynomial blowup and $\Delta = \Omega(\log n)$, such reductions need $B \geq (\log n)^{(1+c_0)/4-o(1)}$.*

The proof turns such a reduction into a Merlin–Arthur protocol for disjointness, in which Merlin names the copy and the players estimate the inner product from random-sign projections, and applies Klauck's lower bound [Kla03]. Taken together, Theorems C and D say that the two readings of the threshold come apart. Measured by the spread, hardness at $(1+\theta)\ln n$ follows, in dimension

Regime	Upper bound (unconditional)	Lower bound, for every $\gamma > 0$ (time $n^{2-\gamma}$ unless noted)
constant precision, $C \log n \le d \le n^{o(1)}$	spread $\kappa \ln n$: time $n^{1+\min\{\kappa,1\}+o(1)}$ (Corollary 3.7)	spread $\kappa \ln n$: time $n^{1+\min\{\kappa,1\}-\gamma}$, LDOV, $d = O(\log n)$ (Corollary 3.7); spread $(1+\theta)\ln n$: LDOV, $d = O(\log n)$ (Theorem 3.3), or CNF-SETH$_{c_0}$ with $c_0 < 2$, $d = n^{o(1)}$ (Theorem 5.3); spread $\ln n + O(1)$: LDOV, $d = (\log n)^{O(1)}$ (Corollary 3.4); spread $d\ln(16n)$ at precision $1/6$: CNF-SETH, $d = (\log n)^{O(1)}$ (Proposition 5.1)
precision $n^{-\beta}$, $d = O(\log n)$	spread $\le \kappa \ln n$ with $\kappa = \kappa(C,\beta,\eta)$: $O(n^{1+\eta})$ operations on $O(\log^2 n)$-bit rationals (Proposition 4.4)	spread $\kappa \ln n$ once $\beta \ge O(\log(2 + 1/\kappa))$: LDOV (Corollary 4.5)
precision about $1/n$, $d \ge (\log n)^{1+\Omega(1)}$	see the next row (Proposition 4.6)	spread $O(\log n/\log\log n)$ at precision $n^{-1-O(1/\log\log n)}$; spread $O(1)$ at $d = O(n^{\delta})$ and precision $\Theta(1/n)$: LDOV (Corollary 4.2)
precision $n^{-\beta}$, $d = (\log n)^C$, $C > 1$	spread $\le (\log n)^{1-a}$ for $a \in (4(\beta+5)(C-1)/\eta, 1]$: $O(n^{1+\eta})$ operations (Proposition 4.6)	spread $O(\log n/\log\log n)$ for $\beta > 1$: LDOV (Corollary 4.2); the gap is Open Problem 1
entry bound B, normalization of [AS23]	$B^2 \le (1-\theta)\ln n/2$ at constant precision (Corollary 3.2, since $W \le 2B^2$)	$B^2 = O(\log n)$: LDOV (Remark B.1). Barrier: gap embeddings from OV in dimension $(\log n)^{c_0}$, $c_0 > 1$, with $p = O(\log n)$ and $\Delta = \Omega(\log n)$ need $B^2 \ge (\log n)^{(1+c_0)/2-o(1)}$ (Corollary 5.7)

Table 1: Fast attention after the refutation of SETH. The spread is the largest difference of two logits in a row; $\theta, \kappa, \eta, \delta, \beta, a > 0$, C, c_0 and the exponent in $(\log n)^{1+\Omega(1)}$ are constants, and the constants in the lower bounds depend on γ and on them.

$n^{o(1)}$, from a hypothesis about satisfiability with clause exponent below 2, and the construction behind it, which builds one gap instance for each proof of a one-message Merlin–Arthur protocol for disjointness (Lemma 5.2), needs clause exponent below 2 for every such protocol (Remark 5.4). Measured by the entry bound, reductions of this kind with polynomial blowup and gap $\Omega(\log n)$ cannot derive $B = \Theta(\sqrt{\log n})$ from OV in dimension $(\log n)^{1+\Omega(1)}$; for them, the entry form of the threshold needs LDOV. Table 1 collects the picture.

1.2 What is new

The ledger of [Li26, Section 6, group (6)] states two facts: the quadratic attention bounds rest on LDOV after the refutation, and bounds proved for every $d = \omega(\log n)$ hold under CNF-SETH in dimension polylog n. This paper turns the first statement into a characterization: at constant precision the exponent is determined at every spread and in every dimension between $\omega(\log n)$ and $n^{o(1)}$, and the constant $\ln n$ of the threshold is exact (Theorem A). It makes the second explicit for attention (Theorem C(i)); under the stronger hypothesis of clause exponent below 2 it obtains the lower bound at spread $(1+\theta)\ln n$ in dimension $n^{o(1)}$ (Theorem C(ii)), and it

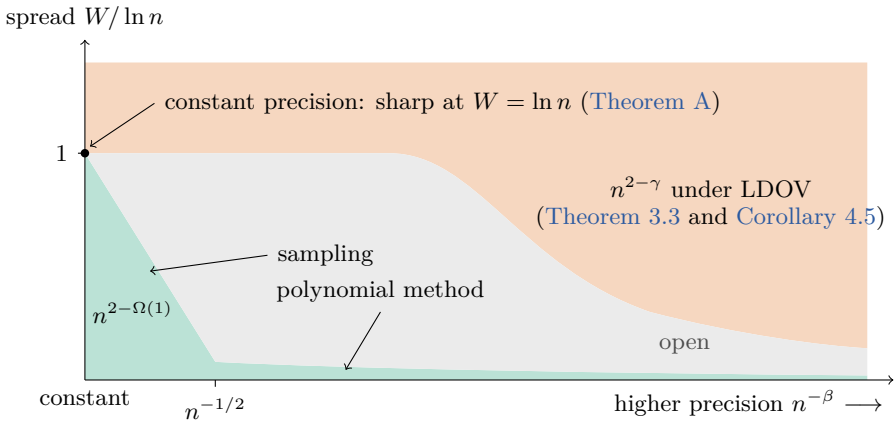


Figure 1: Schematic phase diagram in dimension $O(\log n)$, not to scale; the dimension constant depends on γ, and at constant precision also on θ. At constant precision the easy and hard regions meet at spread $\ln n$, and below it the exponent is $1 + W/\ln n$ (Corollary 3.7). At precision $n^{-\beta}$ and spread $(1 - 2\beta - \xi)\ln n$, for a constant $\xi > 0$, sampling takes time $n^{2-\xi+o(1)}$; beyond $\beta = 1/2$ the polynomial method takes over. At precision $n^{-\beta}$ both boundaries are $\Theta_\beta(\ln n)$ and tend to 0 as β grows. Under CNF-SETH alone we prove hardness at spread $d\ln(16n)$ in dimension $d = (\log n)^{O(1)}$ (Proposition 5.1).

shows that its construction, one gap instance for each proof of a Merlin–Arthur protocol, needs clause exponent below 2 for every such protocol (Remark 5.4). The dependence on precision and dimension (Theorem B) and the barrier (Theorem D) are new. Relative to the attention literature, Theorem A locates the threshold exactly in the spread, at $(1 \pm \theta)\ln n$, and gives the exponent below it, where [AS23, HAPM+26] give hardness at $B = C'\sqrt{\log n}$ with C' unspecified; in the entry form the constant remains open (Section 3.2). Theorem B shows why the threshold of [AS23] has no single constant at polynomial precision, and that dimension, which does not affect the exponent at constant precision between $C\log n$ and $n^{o(1)}$, matters there. Section 6 maps the conditional lower bounds for attention in the literature, most of them SETH-based, to the hypotheses that now support them; two families turn out to rest on refuted hypotheses, one of them in part of its parameter range.

The upper bounds are uniform sampling and the polynomial method of [AS23]. The lower bounds use the distributed PCP of Abboud, Rubinstein and Williams [ARW17] in the efficient form of Rubinstein [Rub18], as packaged by Chen [Che18], the counting argument of [GHS+26] in a general base and over blocks, and Williams' reduction [Wil05]. The barrier uses random-sign projections, Klauck's rectangle argument [Kla03] and Razborov's corruption bound [Raz92].

Organization. Section 2 defines the problem and states the hypotheses and imported theorems. Sections 3 to 5 prove Theorems A and B, and Theorems C and D, with full proofs in Appendices B to D, written as structured proofs in the style of Lamport [Lam95, Lam12]. Appendix A is a roadmap of these proofs, and Figure 2 there shows how the results depend on each other. The subsections of the appendices follow those of the body: Appendix B.3, for example, holds the proofs for Section 3.3. Section 6 is the ledger, and Section 7 discusses related work and open problems.

AI disclosure. This paper was written with Claude Opus 5.5 (Anthropic), working as an agent in Claude Code. The author posed the question and directed the presentation. Claude chose the

results, proved them, wrote the text and drew the figures. Separate Claude agents searched the literature, audited the proofs and reviewed the paper as referees, and the paper was revised to address their findings. At the author's request, ChatGPT (OpenAI) then reviewed the proofs and the tables of prior results and found no fatal or major problems; Claude checked each of its minor findings against the proofs and the cited sources, and corrected the text without changing any proof. The gadgets, the counting identities and the reductions for the precision bounds were tested on small random instances, the identities in exact arithmetic. The author is responsible for the content.

2 Preliminaries

We write log for $\log_2$, ln for the natural logarithm, and $[n] = \{1, \dots, n\}$. For $a, b \in \{0,1\}^D$ the inner product $\langle a, b\rangle$ counts common ones; a and b are *orthogonal* if $\langle a, b\rangle = 0$.

2.1 Attention and its parameters

Definition 2.1 (Attention)**.** An *attention instance* consists of queries $q_1, \dots, q_n \in \mathbb{R}^d$, keys $k_1, \dots, k_{n'} \in \mathbb{R}^d$ and values $v_1, \dots, v_{n'} \in [-1,1]^{d_v}$. The *logits* are $\ell_{ij} = \langle q_i, k_j\rangle$, the *attention weights* are $\pi_{ij} = e^{\ell_{ij}} / \sum_{j'} e^{\ell_{ij'}}$, and the *output* of query i is $o_i = \sum_j \pi_{ij} v_j \in [-1,1]^{d_v}$. The *spread* of row i is $\max_j \ell_{ij} - \min_j \ell_{ij}$, and the spread W of the instance is the largest spread of a row.

Matrix notation writes the output as $D^{-1}AV$ with $A = \exp(QK^\top)$ and $D = \operatorname{diag}(A\mathbf{1})$. Every normalization used in the literature is a rescaling of the queries. Alman and Song [AS23] use $A = \exp(QK^\top/d)$ with all entries of Q, K in $[-B, B]$, so every logit lies in $[-B^2, B^2]$ and $W \le 2B^2$. Gupta, Huang, Saha, Xu and Ye [GHS+26] use $A = \exp(QK^\top)$ with entries in $[-B, B]$, so $W \le 2dB^2$. Transformers use $A = \exp(QK^\top/\sqrt{d})$. We state all results for the spread of the logits that enter the softmax. It does not depend on how a normalization is split between the queries, the keys and a factor such as $1/d$, and it is invariant under adding a constant to all logits of a row.

Definition 2.2 (Approximate attention)**.** $\mathsf{ATT}(n, d, W, \varepsilon)$ is the following problem. The input is an attention instance with n queries, n keys, dimension d, $d_v \le d$ value coordinates, and spread at most W. The goal is to output vectors $\tilde{o}_i$ with $\|\tilde{o}_i - o_i\|_\infty \le \varepsilon$ for all i. A randomized algorithm solves it if it succeeds with probability at least $2/3$.

An instance with fewer queries, or with dimension below d, is an instance of the larger size after padding: zero queries add rows we ignore, and zero coordinates leave every logit unchanged. Inputs are rational numbers of $O(\log(n/\varepsilon))$ bits, and the algorithms of Section 3 use arithmetic operations on numbers of this length, including the exponential function evaluated to relative precision $\varepsilon/100$; the polynomial method of Section 4 uses exact rational arithmetic on numbers of $O(\log^2 n)$ bits. The following stability fact is used throughout.

Lemma 2.3 (Logit perturbation)**.** *Let two instances share their values and have logits ℓ_{ij} and ℓ'_{ij} with $|\ell_{ij} - \ell'_{ij}| \le \eta$ for all j and a fixed row i. Then $\|o_i - o'_i\|_\infty \le e^{2\eta} - 1$.*

2.2 Hypotheses after the refutation

Orthogonal Vectors ($\mathsf{OV}_{n,D}$) asks whether two given lists A and B of n vectors each in $\{0,1\}^D$ contain an orthogonal pair. Repetitions are allowed, and the size of a list counts its entries; all collections of vectors in this paper, such as the instances of Imported Theorems 2.7 and 2.8 and Lemma 5.2, are lists in this sense, and $a \in A$ ranges over the entries of A. Brute force takes time $O(n^2 D)$. For $D = c \log n$ the best algorithm takes time $n^{2-1/O(\log c)}$ [AWY15, CW16]. All hypotheses below are about randomized algorithms with error probability at most $1/3$.

Hypothesis 2.4 (Low-dimensional OV hypothesis, LDOV)**.** For every $\gamma > 0$ there is an integer $c \geq 1$ such that $\mathsf{OV}_{n,\lceil c \log n \rceil}$ cannot be solved in time $O(n^{2-\gamma})$.

Hypothesis 2.5 (CNF-SETH)**.** For every $\gamma > 0$ there is $c_0 \geq 1$ such that satisfiability of CNF formulas with N variables and at most N^{c_0} clauses cannot be decided in time $O(2^{(1-\gamma)N})$.

Hypothesis 2.6 (CNF-SETH with clause exponent c_0)**.** For a constant $c_0 \geq 1$, the hypothesis CNF-SETH$_{c_0}$ states that for every $\gamma > 0$, satisfiability of CNF formulas with N variables and at most N^{c_0} clauses cannot be decided in time $O(2^{(1-\gamma)N})$.

CNF-SETH$_{c_0}$ implies CNF-SETH, and it gets stronger as c_0 decreases. Williams' reduction [Wil05] connects these hypotheses to OV.

Imported Theorem 2.7 (Williams [Wil05])**.** *A CNF formula with N variables, N even, and M clauses can be transformed in time $2^{N/2}\mathrm{poly}(N, M)$ into an instance of $\mathsf{OV}_{n,M}$ with $n = 2^{N/2}$ that has an orthogonal pair if and only if the formula is satisfiable.*

So CNF-SETH implies that for every $\gamma > 0$ there is c_0 such that $\mathsf{OV}_{n,D}$ with $D = \lceil (2 \log n)^{c_0} \rceil$ has no $O(n^{2-\gamma})$ algorithm, and CNF-SETH$_{c_0}$ gives the same with one exponent c_0 for all γ. Before the refutation, SETH implied Hypothesis 2.4 through the sparsification lemma [IPZ01], which reduces k-SAT to k-SAT with linearly many clauses. This route is closed: CNF-SAT with $O(N)$ clauses of arbitrary width is solvable in time $O(2^{\varepsilon N})$ for every $\varepsilon > 0$ [Li26, Proposition 6.1], so CNF-SETH$_1$ is false, and no hypothesis about satisfiability that is still open is known to imply Hypothesis 2.4. CNF-SETH$_{c_0}$ is open for every $c_0 > 1$. The same applies to the k-party version: SETH implied that k-OV in dimension $c \log n$ needs time $n^{k-o(1)}$, while CNF-SETH, by splitting the variables into k parts, gives this in dimension $(\log n)^{O(1)}$.

2.3 Imported theorems

Three results are used as black boxes. The first packages the distributed PCP of Abboud, Rubinstein and Williams [ARW17], made efficient by Rubinstein [Rub18] with algebraic-geometry codes, as stated by Chen [Che18]. The *Boolean Max-IP* problem asks for the largest inner product between a vector of one list and a vector of another, each of n vectors in $\{0,1\}^d$.

Imported Theorem 2.8 ([Che18, Lemma 3.3 and its proof in Appendix D])**.** *For every integer $c \geq 1$ and real $\varepsilon \in (0, 1]$ there is n_0 such that for all $n \geq n_0$ and every real $\tau \in [4, n]$, an instance (A, B) of $\mathsf{OV}_{n,\lceil c \log n \rceil}$ can be mapped to at most $2n^{\varepsilon}$ instances (A_ψ, B_ψ) of Boolean Max-IP, each with n vectors per side in a dimension $d_{\mathrm{C}} \leq \tau^{\mathrm{poly}(c/\varepsilon)} \log n$, and to an integer T with $c \log n \leq T \leq d_{\mathrm{C}}$, such that:*

- *(i) every inner product between a vector of A_ψ and a vector of B_ψ lies in $[0, T]$;*
- *(ii) if (A, B) has an orthogonal pair, then some instance has a pair with inner product T;*
- *(iii) otherwise every pair in every instance has inner product at most T/τ;*
- *(iv) the map works vector by vector: there are maps $x(\psi, \cdot)$ and y, which depend only on n, c, ε and τ, with $A_\psi = (x(\psi, a))_{a \in A}$ and $B_\psi = (y(b))_{b \in B}$, and for all $a, b \in \{0,1\}^{\lceil c \log n \rceil}$, $\langle x(\psi, a), y(b) \rangle = T$ for some ψ if $\langle a, b \rangle = 0$, and $\langle x(\psi, a), y(b) \rangle \leq T/\tau$ for all ψ if $\langle a, b \rangle \geq 1$.*

The instances can be computed in time $n^{1+\varepsilon} d_{\mathrm{C}} \mathrm{poly}(\log n)$.

Items (i) and (iv), the integer T and the running time are read off the proof. There, $T = 2^r$ is the number of random strings of an amplified protocol, and ψ ranges over Merlin's proofs. The vector $y(b)$ has exactly one coordinate equal to 1 in each of the T blocks indexed by the random strings, at

Bob's message for b, which depends only on b and the random string [Che18, Definition 2.6]; so $T \le d_{\mathrm{C}}$. The coordinates of $x(\psi, a)$ record whether Alice, holding a and the proof ψ, accepts a given message on a given random string, so $\langle x(\psi,a), y(b)\rangle$ is the number of random strings on which Alice accepts, and (iv) is the completeness and soundness of the protocol. Every coordinate is a predicate of the computationally efficient protocol of [Rub18]. The base protocol is the unrepeated one inside Rubinstein's proof of [Rub18, Theorem 3.1]: with s a constant depending on c/ε, it uses $\log m + O(1)$ coins, has perfect completeness and soundness at most $9/10$. An expander walk of length $O(\log \tau)$ brings the soundness to $1/\tau$ with $\log m + O(\log \tau)$ coins; Chen [Che18, Lemma D.2] carries this out from soundness $1/2$, and the hitting property of expander walks gives the same from soundness $9/10$. Padding the protocol with $O(1)$ unused coins makes $T \ge m$ and multiplies the dimension by a constant. This is why $T = O(\tau^{O(1)} \log n)$ and the dimension stays $O(\log n)$ for constant τ. Chen states the lemma for every $\tau \ge 2$ and applies it with τ a function of n [Che18, Corollary 3.4]; in the construction, n_0 comes only from the parameters of Rubinstein's protocol, which depend on c and ε but not on τ. For dimensions $D = \omega(\log n)$ we use the underlying protocol directly. In a *one-message MA protocol* for disjointness, Merlin sends a proof to Alice, Alice and Bob share uniformly random coins, Bob sends one message to Alice, and Alice accepts or rejects.

Imported Theorem 2.9 ([Rub18, Theorem 3.1])**.** *There are absolute constants a_1, a_2, a_3 such that for all integers m and $s \in [2, m]$ there is a computationally efficient one-message MA protocol for disjointness of $X, Y \subseteq [m]$ in which Merlin sends Alice at most $a_1(m/s)\log s$ bits, Alice and Bob share at most $a_2 \log m$ uniform coins, and Bob sends Alice at most $a_3 s \log s$ bits that depend only on Y and the coins. If $X \cap Y = \emptyset$, some proof makes Alice accept for all coins. If $X \cap Y \neq \emptyset$, Alice accepts with probability at most $1/2$ for every proof.*

The third result is Razborov's corruption bound for disjointness. Write $\mathsf{DISJ}_D(x, y) = 1$ if $x, y \in \{0,1\}^D$ are disjoint, that is, orthogonal.

Imported Theorem 2.10 (Razborov [Raz92])**.** *There are absolute constants $\alpha_1, \varepsilon_1 > 0$ and, for every $D \ge 1$, a distribution μ on $\{0,1\}^D \times \{0,1\}^D$ with $\mu(\mathsf{DISJ}_D^{-1}(1)) = 3/4$ such that every rectangle $X \times Y$ satisfies $\mu((X \times Y) \cap \mathsf{DISJ}_D^{-1}(0)) \ge \alpha_1 \, \mu((X \times Y) \cap \mathsf{DISJ}_D^{-1}(1)) - 2^{-\varepsilon_1 D}$.*

Razborov proves this for $D = 4\ell - 1$; for other D, place the distribution on the first $4\ell - 1 \ge D - 4$ coordinates and decrease ε_1, and for $D \le 6$ take any distribution with the stated mass and a small ε_1. The direction matters: rectangles that are mostly disjoint must be small. The converse fails, since $\{x \ni i\} \times \{y \ni i\}$ contains only intersecting pairs.

3 The concentration threshold

This section proves Theorem A. The upper bound is uniform sampling; its cost is governed by how much weight a single key can receive. The lower bounds show that under Hypothesis 2.4 this cost cannot be improved by a polynomial factor at any spread, and that once the spread allows one key to outweigh all others together, constant precision already needs quadratic time.

3.1 Sampling

Proposition 3.1 (Sampling)**.** *Consider an attention instance with n queries, n' keys, dimension d and d_v value coordinates. Let $F \ge 1$ satisfy $F \ge n'\pi_{ij}$ for all i, j, and let $\varepsilon, \delta \in (0, 1]$. There is a randomized algorithm that, with probability at least $1 - \delta$, outputs $\tilde{o}_i$ with $\|\tilde{o}_i - o_i\|_\infty \le \varepsilon$ for all i, using $O(nm(d + d_v))$ operations with*

$$m = \left\lceil 40\, F \ln(2n(1 + d_v)/\delta)/\varepsilon^2 \right\rceil.$$

Every instance with spread at most W satisfies the hypothesis with $F = e^W$.

The algorithm samples m keys uniformly with replacement for each query and returns the self-normalized average $\sum_s e^{\ell_{ijs}} v_{j_s} / \sum_s e^{\ell_{ijs}}$. Bernstein's inequality controls numerator and denominator, because the variance of a weight is at most its mean times its maximum (Appendix B.1). The quantity $n' \max_j \pi_{ij}$ is the factor by which the largest attention weight exceeds the uniform weight $1/n'$, so the cost is n^2 times the largest attention weight, up to the factor $(d + d_v) \log n / \varepsilon^2$.

Corollary 3.2. *For every $\kappa \in [0, 1]$ and $\varepsilon \in (0, 1]$, $\mathsf{ATT}(n, d, \kappa \ln n, \varepsilon)$ can be solved in time $O(n^{1+\kappa} d \log(nd)/\varepsilon^2)$. The same holds for instances in which every attention weight is at most $n^{\kappa-1}$.*

Proof. Apply Proposition 3.1 with $n' = n$, $\delta = 1/3$, $d_v \le d$ and $F = e^{\kappa \ln n} = n^\kappa$, respectively $F = n \cdot n^{\kappa-1}$. □

Sampling keys is the idea behind several practical attention approximations, such as KDEformer [ZHDK23] and HyperAttention [HJK+24], whose guarantees are stated in terms of column norms of the attention matrix and stable ranks. Proposition 3.1 is the simplest member of this family, and under Hypothesis 2.4, Corollary 3.7 below shows that its exponent is optimal in terms of the spread.

3.2 Quadratic time above the threshold

Theorem 3.3 (Hardness above the threshold)**.** *Assume Hypothesis 2.4. For all $\gamma \in (0, 1)$ and $\theta \in (0, 1]$ there is a constant $C \ge 1$ such that for every $\varepsilon \in (0, 1/2)$, $\mathsf{ATT}(n, \lceil C \log n \rceil, (1 + \theta) \ln n, \varepsilon)$ cannot be solved in time $O(n^{2-\gamma})$. This holds already for instances with $d_v = 1$, values in $\{0, 1\}$, and all logits in $[0, (1 + \theta) \ln n]$.*

For values in $\{0, 1\}$, precision $1/2$ is trivial: output $1/2$ everywhere. For such values, Theorem 3.3 therefore covers every nontrivial constant precision. The proof gives a constant C that is polynomial in $1/\theta$, with a degree that depends on γ and on the constant of Hypothesis 2.4; so in a fixed dimension $C \log n$ the hardness holds at spread $(1 + C^{-\Omega_\gamma(1)}) \ln n$.

Proof idea. Take the gap instances of Imported Theorem 2.8 with $\tau \ge 4/\theta$: every inner product lies in $[0, T]$, a YES instance has a pair at T, and a NO instance has all pairs at most T/τ. Let $\eta = (1 - 2\varepsilon)/4$. Scale the vectors by $\lambda \approx (\ln n + 2 \ln(1/\eta))/((1 - 1/\tau)T)$ and add one dummy key whose logit sits $\ln((1 - \eta)/\eta)$ below the top value λT. If some key of a row is at λT, the real keys of that row carry at least $1 - \eta$ of its weight. In a NO instance they carry at most η, because each lies $\lambda T(1 - 1/\tau)$ below the top. All logits lie in $[0, \lambda T]$, and $\lambda T \le (1 + \theta) \ln n$ for large n. Each of the at most $2n^{\gamma/4}$ instances takes $O(\log n)$ attention calls, so an $O(n^{2-\gamma})$ algorithm gives an $O(n^{2-\gamma/2})$ algorithm for OV. Appendix B.2 gives the details. □

Imported Theorem 2.8 allows the gap τ to grow with n. With $\tau = \lceil \log n \rceil$ the same proof gives an additive version, at the price of polylogarithmic dimension.

Corollary 3.4 (Additive version)**.** *Assume Hypothesis 2.4. For all $\gamma \in (0, 1)$ and $\varepsilon \in (0, 1/2)$ there is $C \ge 1$ such that $\mathsf{ATT}(n, \lceil (\log n)^C \rceil, \ln n + 2 \ln \frac{4}{1-2\varepsilon} + 2, \varepsilon)$ cannot be solved in time $O(n^{2-\gamma})$. For $\varepsilon = 1/6$ the spread is below $\ln n + 6$. This holds already for instances with $d_v = 1$ and values in $\{0, 1\}$.*

Remark 3.5 (Values in $\{-1,1\}$). Replacing every value v by $2v-1$ maps an attention instance with values in $\{0,1\}$ to an instance with the same logits and values in $\{-1,1\}$, and every output o to $2o-1$. So an algorithm with precision ε on the new instances gives one with precision $\varepsilon/2$ on the old ones, at the cost of $O(n)$ further operations when $d_v = 1$. Hence Theorem 3.3, and Theorems 3.6 and 5.3 below, hold for every $\varepsilon \in (0,1)$ on instances with values in $\{-1,1\}$, and so does Corollary 3.4 with spread $\ln n + 2\ln\frac{4}{1-\varepsilon} + 2$. With values in $[-1,1]$, precision 1 is trivial: output 0 everywhere. So these results cover every nontrivial constant precision.

3.3 The exponent below the threshold

Below the threshold, splitting the keys into groups gives a matching lower bound at every spread.

Theorem 3.6 (Hardness below the threshold)**.** *Assume Hypothesis 2.4. For all $\alpha \in (0,1]$, $\gamma \in (0,\alpha)$ and $\theta \in (0,1]$ there is a constant $C \geq 1$ such that for every $\varepsilon \in (0,1/2)$, $\mathsf{ATT}(n, \lceil C\log n\rceil, (1+\theta)\alpha\ln n, \varepsilon)$ cannot be solved in time $O(n^{1+\alpha-\gamma})$. This holds already for instances with $d_v = 1$, values in $\{0,1\}$, and all logits in $[0,(1+\theta)\alpha\ln n]$.*

Proof idea. Under Hypothesis 2.4, OV with n vectors on one side and n^α on the other needs time $n^{1+\alpha-o(1)}$, because splitting one side of a balanced instance into $n^{1-\alpha}$ groups gives $n^{1-\alpha}$ such instances. The gadget of Theorem 3.3, applied to one group through the vector-by-vector form of Imported Theorem 2.8, has n^α real keys, so a top logit $\alpha\ln n + O(1)$ above the others suffices. Copying every key, the dummy key included, about $n^{1-\alpha}$ times multiplies every term of the numerator and of the denominator of the softmax by the same factor, so it leaves the outputs and the spread unchanged and brings the number of keys to between n and $2n$. An algorithm faster than $n^{1+\alpha-\gamma}$ on these instances would solve all groups in time $n^{2-\Omega(\gamma)}$. Appendix B.3 gives the details. □

Corollary 3.7 (The exponent)**.** *Fix $\varepsilon \in (0,1)$ and $\kappa > 0$. Then $\mathsf{ATT}(n,d,\kappa\ln n,\varepsilon)$ can be solved in time $n^{1+\min\{\kappa,1\}+o(1)}$ for every $d \leq n^{o(1)}$. Under Hypothesis 2.4, for every $\gamma > 0$ there is C such that $\mathsf{ATT}(n,d,\kappa\ln n,\varepsilon)$ has no $O(n^{1+\min\{\kappa,1\}-\gamma})$ algorithm for any d with $\lceil C\log n\rceil \leq d \leq n^{o(1)}$. This holds already for instances with $d_v = 1$ and values in $\{-1,1\}$, and with values in $\{0,1\}$ if $\varepsilon < 1/2$.*

Proof. Let $\kappa' = \min\{\kappa,1\}$. For $\kappa < 1$ the first part is Corollary 3.2; for $\kappa \geq 1$ direct computation takes $O(n^2 d)$ operations, with the exponentials evaluated as in step ⟨1⟩6 of the proof of Proposition 3.1. For the second part it suffices to treat $\gamma < \kappa'/2$: for larger γ, the constant C for $\kappa'/4$ works, since an $O(n^{1+\kappa'-\gamma})$ algorithm is an $O(n^{1+\kappa'-\kappa'/4})$ algorithm. Choose $\theta \in (0,1]$ with $\kappa'/(1+\theta) \geq \kappa' - \gamma/2$ and let $\alpha = \kappa'/(1+\theta)$. Then $\gamma/2 < \kappa'/4 < \alpha$, and for $\varepsilon < 1/2$, Theorem 3.6 for α, $\gamma/2$ and θ gives C such that spread $(1+\theta)\alpha\ln n = \kappa'\ln n$ has no $O(n^{1+\alpha-\gamma/2})$ algorithm in dimension $\lceil C\log n\rceil$, already for $d_v = 1$ and values in $\{0,1\}$. Since $1+\alpha-\gamma/2 \geq 1+\kappa'-\gamma$ and $\kappa' \leq \kappa$, this proves the claim in that dimension. For every $\varepsilon \in (0,1)$ and values in $\{-1,1\}$, the same C works by Remark 3.5 at precision $\varepsilon/2$; the further $O(n)$ operations are absorbed, since $1+\kappa'-\gamma > 1$. Padding with zero coordinates preserves logits, and for $d \leq n^{o(1)}$ writing the padded instance takes $n^{1+o(1)}$ operations, which are absorbed too. So the bound holds for every d with $\lceil C\log n\rceil \leq d \leq n^{o(1)}$. □

So under Hypothesis 2.4, at every nontrivial constant precision and in every dimension between $\omega(\log n)$ and $n^{o(1)}$, the time exponent of attention with spread $\kappa\ln n$ is exactly $1+\min\{\kappa,1\}$: up to factors $n^{o(1)}$ the cost is $n\cdot e^W$ until it reaches n^2, and uniform sampling is optimal. The transition to quadratic time sits at $W = \ln n$, the spread at which a single key can first outweigh all others together. At spread $(1-\theta)\ln n$ every attention weight is at most $e^W/n = n^{-\theta}$, and sampling sees enough keys. At spread $(1+\theta)\ln n$ a row can put almost all of its weight on the keys at the top

logit, and deciding whether such a row exists is as hard as OV. For each $\gamma > 0$ the lower bounds hold in dimension $C \log n$ for a constant C that depends on κ and γ; in polylogarithmic dimension, Corollary 3.4 reaches spread $\ln n + O(1)$. Below dimension $C \log n$ the dimension matters: for constant d, Gupta et al. [GHS+26] compute attention in time $\tilde{O}(n^{2-1/d})$ at every spread.

The spread is a worst-case parameter. A single key with a large logit in every row, such as an attention sink, makes the spread large while the other keys remain cheap to sample, and methods such as HyperAttention [HJK+24] find and remove such keys first. In the instances of Theorem 3.3 the difficulty is to decide whether some row has a key at the top logit; Haverbeck et al. [HAPM+26, Theorem 5.1] show directly that finding the keys that carry most of the weight is hard.

Comparison with the bounded-entry threshold. In the normalization of Alman and Song, Corollary 3.2 applies whenever $B \le \sqrt{(1-\theta)\ln n/2}$, because $W \le 2B^2$. Theorem 3.3 gives instances with $B = O_{\gamma,\theta,\varepsilon}(\sqrt{\log n})$ (Remark B.1), so the threshold $B = \Theta(\sqrt{\log n})$ of [AS23] holds at constant precision, under Hypothesis 2.4. In terms of B the constant remains open: $2B^2$ bounds the spread only up to a factor that depends on how many coordinates of the vectors are nonzero. In terms of the spread it is exact. Haverbeck et al. [HAPM+26, Theorem 3.7] prove hardness at every nontrivial constant precision (additive error below $1/2$ for values in $[0,1]$, relative error below 1) with $d = C_d \log n$ and $B = C_B\sqrt{\log n}$, for constants C_d, C_B that depend on the precision and are not specified; their reduction also goes through Rubinstein's gap problem, so it rests on Hypothesis 2.4 as well. Gupta et al. [GHS+26, Theorem 5.7] give hardness at $d = \Theta(\log n)$ with entries $\ln 2$ in the unnormalized form, that is, with spread $\Theta(\log n)$ and polynomial precision.

4 Precision and dimension move the threshold

The threshold $W = \ln n$ is a statement about constant precision. Precision $1/\mathrm{poly}(n)$, as in [AS23], keeps errors small when attention layers are composed. At that precision the picture changes: there is no single constant, higher precision makes smaller spreads hard, and, unlike at constant precision, the dimension matters. This section proves Theorem B. We write the precision as $n^{-\beta}$.

4.1 Counting modulo $r+1$

Gupta et al. [GHS+26, Theorem 5.7 and Lemma 5.8] observed that attention with logits $(\ln 2)\langle a, b\rangle$ and precision $n^{-O(1)}$ reveals the parity of the number of orthogonal partners of a. The point is that $\sum_b 2^{\langle a,b\rangle}$ is an integer whose parity is the parity of the number of b with $\langle a, b\rangle = 0$. We make two changes. The base $1 + 1/r$ instead of 2 makes the logits $\ln(1+1/r)\langle a, b\rangle$ small, while $r^m \sum_b (1+1/r)^{\langle a,b\rangle}$ is still an integer that determines the number of orthogonal partners modulo $r+1$. And counting intersecting *blocks* instead of intersecting coordinates reduces the number of digits that must be resolved, at the price of dimension. For a partition of $[m]$ into K blocks, let

$$F(a,b) = \big|\{t \le K : \ a \text{ and } b \text{ share a one in block } t\}\big|.$$

Then $F(a,b) = 0$ exactly when $a \perp b$, and $F(a,b) = \langle \phi(a), \psi(b)\rangle$ for vectors of dimension $\sum_t (2^{|B_t|}-1)$, by inclusion–exclusion inside each block (Appendix C.1).

Theorem 4.1 (Precision lower bound)**.** *Assume Hypothesis 2.4. For every $\gamma \in (0,1)$ there is an integer $c \ge 1$ with the following property. Let $r, w : \mathbb{N} \to \mathbb{N}$ be computable in time $O(n)$, with $r(n), w(n) \ge 1$ for all n, and $r(n) \le n$ and $w(n) \le \frac{1}{2}\log n$ for all large n. Let*

$$K = \lceil \lceil c \log n \rceil / w \rceil, \qquad d = K(2^w - 1), \qquad W = K \ln(1 + 1/r), \qquad \varepsilon = \frac{1}{64\, n\, (r+1)^{2K}},$$

with $r = r(n)$ and $w = w(n)$. Then $\mathsf{ATT}(n, d, W, \varepsilon)$ cannot be solved in time $O(n^{2-\gamma})$. This holds already for $d_v = 1$, values in $\{0, 1\}$, and all logits in $[0, W]$.

Here w is the size of the blocks. With $w = 1$ the blocks are single coordinates, $F(a, b) = \langle a, b\rangle$, and the dimension is $K = \lceil c \log n\rceil$.

Proof idea. Fix $a \in A$, keep each $b \in B$ with probability $1/2$, and replace the dropped vectors by the all-ones vector, which shares a one with every $a \neq 0$. The queries are $\lambda\phi(a)$ with $\lambda \approx \ln(1+1/r)$, the keys are ψ of the kept vectors, and n further keys equal to 0 carry value 1. The output of a is $n/(n + S_a)$ with $S_a = \sum_b (1 + 1/r)^{F(a,b)}$. If the vectors of the OV instance have $K_0 \leq K$ blocks, precision ε recovers the integer $r^{K_0} S_a$ exactly. Modulo $r + 1$, every term with $F \geq 1$ vanishes, so $r^{K_0} S_a \equiv r^{K_0} N_a$, where N_a is the number of kept orthogonal partners of a, and r^{K_0} is invertible modulo $r + 1$. If a has an orthogonal partner, then $N_a \not\equiv 0 \pmod{r+1}$ with probability at least $1/2$. Four rounds detect an orthogonal pair with probability $15/16$. Appendix C.1 gives the details. □

Three choices of the parameters give the following.

Corollary 4.2. *Assume Hypothesis 2.4, and let $\gamma \in (0, 1)$ and c be as in Theorem 4.1. None of the following can be solved in time $O(n^{2-\gamma})$.*

(a) Dimension $\lceil c \log n\rceil$. *For every $\kappa > 0$: spread $\kappa \ln n$ at precision*

$$\frac{1}{64\,(c_\kappa + 3)^2\, n^{1+2c\log(c_\kappa+3)}}, \qquad c_\kappa = \frac{c}{\kappa \ln 2};$$

and spread 1 at precision $n^{-O(c \log(c \log n))}$.

(b) Polylogarithmic dimension. *For every rational $C' > 0$: dimension $(\log n)^{1+C'}$, spread at most $\frac{c \ln 2}{C'} \cdot \frac{\log n}{\log\log n} + 1$, and precision $n^{-1-\zeta}$ with $\zeta = O(c/(C' \log\log n))$.*

(c) Polynomial dimension. *For every rational $\delta \in (0, 1/2]$: dimension $(c/\delta + 2)n^\delta$, spread $(c/\delta + 2) \ln 2$, precision $1/(64 \cdot 4^{c/\delta+2} n)$.*

Proof. All statements are for large n. Where the formulas below give r or w a value below 1, which happens only for small n, we use 1 instead; then $r(n) \leq n$ and $w(n) \leq \frac{1}{2} \log n$ hold for all large n. For (a), take $w = 1$, so $K = \lceil c \log n\rceil \leq c \log n + 1$ and $\varepsilon \geq 1/(64(r+1)^2 n^{1+2c\log(r+1)})$. With the constant $r = \lceil c_\kappa\rceil + 1$, the spread is at most $(c \log n + 1)/r \leq \kappa \ln n$ for $n \geq e^{1/\kappa}$, because $r\kappa \ln n \geq c \log n + \kappa \ln n$, and $r + 1 \leq c_\kappa + 3$. With $r(n) = \lceil c \log n\rceil$, the spread is at most $K/r = 1$ and $\varepsilon = n^{-O(c \log(c \log n))}$. For (b), take $r = 1$ and $w = \lceil C' \log\log n\rceil$: then $K \leq (c \log n + 1)/(C' \log\log n) + 1$, the spread is $K \ln 2$, the dimension is at most $2K(\log n)^{C'} \leq (\log n)^{1+C'}$, and $\varepsilon = 1/(64n4^K)$. For (c), take $r = 1$ and $w = \lfloor \delta \log n\rfloor$: then $K \leq c/\delta + 2$, the dimension is at most $K\, n^\delta$, and $\varepsilon = 1/(64n4^K)$. □

Remark 4.3 (Bounded entries). In the normalization of [AS23], case (a) with $r(n) = \lceil c \log n\rceil$ has bounded entries. There $w = 1$, $\phi(a) = a$ and $\psi(b) = b$, so an instance of size n in the proof of Theorem 4.1 has dimension $d = K = \lceil c \log n\rceil = r$, queries $\lambda' a$ with $0 < \lambda' \leq \ln(1 + 1/r) \leq 1/d$, and keys in $\{0, 1\}^d$ after padding. Rescaling the queries by a rational $s \in [\sqrt{d/\lambda'}, 2\sqrt{d/\lambda'}]$ and the keys by d/s, as in Remark B.1, writes the same logits as $\langle q, k\rangle/d$ with all entries bounded by $2\sqrt{d\lambda'} \leq 2$. So entries bounded by 2 are already hard at precision $n^{-O(c \log(c \log n))}$.

4.2 The polynomial method

On the algorithmic side, the polynomial method of Alman and Song [AS23] replaces e^x by a polynomial on the range of the logits. For dimension $O(\log n)$ and precision $n^{-\beta}$ it handles every spread below a constant multiple of $\ln n$, with a constant that tends to 0 as β grows.

Proposition 4.4 (Polynomial method)**.** *For all $C \geq 1$, $\beta \geq 1$ and $\eta \in (0,1)$ there is $\kappa = \kappa(C,\beta,\eta) > 0$ such that the following holds for all sufficiently large n. Every attention instance with n queries, n keys, dimension $d \leq C \log n$, $d_v \leq d$ value coordinates and spread at most $\kappa \ln n$ can be solved to precision $n^{-\beta}$ deterministically with $O(n^{1+\eta})$ arithmetic operations on rationals of $O(\log^2 n)$ bits.*

This is the method of [AS23] with the dependence on β and on the spread made explicit. The proof first shifts each row by its first logit, through one extra coordinate equal to $-\langle q_i, k_1\rangle$ in q_i and to 1 in every key, so that all logits lie in $[-\kappa \ln n, \kappa \ln n]$, and rounds the entries to a grid of $O(\log n)$ bits; then a Taylor polynomial of degree $g = \Theta(\beta \log n / \log(\beta/\kappa))$ has rank $\binom{d+g}{g} = n^{O(\eta)}$ (Appendix C.2).

Corollary 4.5 (Precision decides the threshold)**.** *Assume Hypothesis 2.4. For every $\gamma \in (0,1)$ there is C_γ such that the following holds for instances of dimension at most $C_\gamma \log n$. For every $\beta \geq 1$ there is $\kappa_{\mathrm{easy}}(\beta) > 0$ such that spread $\kappa_{\mathrm{easy}}(\beta) \ln n$ at precision $n^{-\beta}$ can be solved in time $O(n^{2-\gamma})$. For every $\kappa > 0$ there is $\beta_{\mathrm{hard}}(\kappa) = O(C_\gamma \log(2 + C_\gamma/\kappa))$ such that spread $\kappa \ln n$ at precision $n^{-\beta_{\mathrm{hard}}(\kappa)}$ cannot. So the largest spread, in units of $\ln n$, that can be handled in time $O(n^{2-\gamma})$ at precision $n^{-\beta}$ is positive for every β and tends to 0 as $\beta \to \infty$.*

Proof. Let c be the integer of Theorem 4.1 for γ, and $C_\gamma = c + 1$. The first claim is Proposition 4.4 with $C = C_\gamma$ and $\eta = (1-\gamma)/2$, since $O(n^{1+\eta})$ operations on $O(\log^2 n)$-bit numbers take time $O(n^{2-\gamma})$. The second is Corollary 4.2(a), whose precision is $n^{-O(c \log(2+c/\kappa))}$. Every fixed $\kappa > 0$ is therefore hard at precision $n^{-\beta_{\mathrm{hard}}(\kappa)}$, so the largest solvable spread at precision $n^{-\beta}$ is below $\kappa \ln n$ once $\beta \geq \beta_{\mathrm{hard}}(\kappa)$. □

4.3 Dimension matters at high precision

At constant precision the dimension does not affect the exponent once it is at least $C \log n$: Proposition 3.1 works in every dimension, and Theorems 3.3 and 3.6 need only $C \log n$. At precision about $1/n$ it matters a great deal. In dimension $O(\log n)$ and at precision $n^{-1-O(1/\log\log n)}$, the boundary between easy and hard spreads is $\Theta(\log n)$: spread $\kappa \ln n$ is easy for some $\kappa > 0$ by Proposition 4.4 with $\beta = 2$, and spread $(1+\theta)\ln n$ is hard by Theorem 3.3, because hardness at constant precision carries over to every finer precision. By Corollary 4.2, the hard spread falls to $O(\log n / \log\log n)$ in dimension $(\log n)^{1+C'}$ for every $C' > 0$, and to $O(1)$ in dimension n^δ. The mechanism is a trade-off between three parameters of Theorem 4.1: blocks of size w divide the number K of digits, and with it the spread $K \ln(1+1/r)$, by w, and take the w-th root of the excess $(r+1)^{2K}$ in the precision, while the dimension $K(2^w - 1)$ grows exponentially in w.

The upper bound of Proposition 4.4 degrades with the dimension as well.

Proposition 4.6 (Polynomial method in higher dimension)**.** *Let $C > 1$, $\beta \geq 1$ and $\eta \in (0,1)$ with $4(\beta+5)(C-1) < \eta$, and let $a \in (4(\beta+5)(C-1)/\eta, 1]$. For all sufficiently large n, every attention instance with n queries, n keys, dimension $d \leq (\log n)^C$, $d_v \leq d$ value coordinates and spread at most $(\log n)^{1-a}$ can be solved to precision $n^{-\beta}$ deterministically with $O(n^{1+\eta})$ arithmetic operations on rationals of $O(\log^2 n)$ bits.*

The proof (Appendix C.3) is that of Proposition 4.4 with a Taylor polynomial of degree $g = \lceil 2(\beta+5)\ln n/(a \ln\ln n)\rceil$, which has $n^{(1+o(1))2(\beta+5)(C-1)/a}$ monomials in dimension $(\log n)^C$. So in dimension $(\log n)^C$ with $C < 1 + \eta/(4(\beta+5))$, the method handles spread $(\log n)^{1-a}$ for every $a \in (4(\beta+5)(C-1)/\eta, 1]$, below the spread $O(\log n/\log\log n)$ that Corollary 4.2(b) makes hard at every precision $n^{-\beta}$ with $\beta > 1$. For $C \geq 1 + \eta/(4(\beta+5))$ the proposition gives nothing. Open Problem 1 asks where the threshold lies. Padding with zero coordinates preserves logits, so every lower bound transfers to every larger dimension d for which nd stays below its time bound.

5 What satisfiability still supports

Theorems 3.3, 3.6 and 4.1 rest on Hypothesis 2.4, which no open hypothesis about satisfiability is known to imply. This section asks what CNF-SETH gives instead. The answer depends on the clause exponent of the hard formulas and on whether the threshold is measured by the spread or by the entry bound B.

5.1 CNF-SETH in polylogarithmic dimension

Williams' reduction turns CNF-SETH into hardness of OV in dimension $(\log n)^{O(1)}$. Without a gap, the softmax must separate one orthogonal pair from n pairs whose inner products are only one larger, so every unit of inner product costs $\ln n$ in the logit.

Proposition 5.1 (CNF-SETH, exact embedding)**.** *Assume Hypothesis 2.5. For every $\gamma \in (0,1)$ there is $C \geq 1$ such that, with $d = \lceil(\log n)^C\rceil$, the problem $\mathsf{ATT}(n, d, d\ln(16n), 1/6)$ cannot be solved in time $O(n^{2-\gamma})$. This holds already for $d_v = 1$ and values in $\{0,1\}$.*

The queries are $(-\lambda a, -11/10)$ with $\lambda = \lceil 10\ln(14n)\rceil/10$, the keys are $(b, 0)$ with value 1, and a dummy key $(0,1)$ has value 0; the proof is in Appendix D.1. This is the exact embedding of Gupta et al. [GHS+26, Lemma 5.6] combined with Williams' reduction. The reductions of Keles, Wijewardena and Hegde [KWH23], proved for every $d = \omega(\log n)$, are exact as well and survive under CNF-SETH in dimension $(\log n)^{O(1)}$, with the precision they need there (Table 3). In terms of the entry bound of [AS23], the instances of Proposition 5.1 can be written with $B = O(\sqrt{d\log n})$.

5.2 Clause exponent below two

The gap machinery of Imported Theorem 2.8 needs OV in dimension $O(\log n)$. Its engine, Rubinstein's protocol (Imported Theorem 2.9), works in every dimension m, but the instance grows by the factor 2^p for a proof of $p \approx (m/s)\log s$ bits, and the dimension by the factor $2^{O(s\log s)}$ for Bob's message. With $m = (2\log n)^{c_0}$ from Williams' reduction and $s \approx m\log m/(\gamma\log n)$, the first factor is at most $n^{\gamma/8}$, and the second is then $n^{o(1)}$ exactly when $c_0 < 2$.

Lemma 5.2 (Gap instances from MA protocols)**.** *Let $m \geq s \geq 2$ and $k \geq 1$ be integers, and let p, ρ and ℓ be the lengths of Merlin's proof, of the coins and of Bob's message in the protocol of Imported Theorem 2.9 for (m,s). There is a map from instances (A,B) of $\mathsf{OV}_{n,m}$ to 2^p instances of Boolean Max-IP, each with n vectors per side in dimension $2^{k(\rho+\ell)}$, such that with $T_k = 2^{k\rho}$:*

(i) every inner product lies in $[0, T_k]$;
(ii) if (A,B) has an orthogonal pair, some instance has a pair with inner product T_k;
(iii) otherwise every inner product is at most $2^{-k}T_k$.

The map takes time $2^{p+k(\rho+\ell)}\, nk\mathrm{poly}(m)$.

Theorem 5.3 (Clause exponent below two)**.** *Assume* CNF-SETH$_{c_0}$ *for some* $c_0 < 2$*. For all* $\gamma \in (0,1)$*,* $\theta \in (0,1]$ *and* $\varepsilon \in (0,1/2)$ *there is* $K \geq 1$ *such that* $\mathsf{ATT}(n, d(n), (1+\theta)\ln n, \varepsilon)$ *with* $d(n) = \lceil 2^{K(\log n)^{c_0-1}(\log\log n)^2} \rceil = n^{o(1)}$ *cannot be solved in time* $O(n^{2-\gamma})$*. This holds already for instances with* $d_v = 1$*, values in* $\{0,1\}$*, and all logits in* $[0,(1+\theta)\ln n]$*.*

So the lower bound of Theorem 3.3 at spread $(1+\theta)\ln n$ follows, in dimension $n^{o(1)}$ instead of $O(\log n)$, from a hypothesis about satisfiability, provided the hard CNF formulas have N^{c_0} clauses for some $c_0 < 2$. Formulas with $O(N)$ clauses are easy [Li26, Proposition 6.1], so the relevant range is $1 < c_0 < 2$, where the question is open: [Li26, Proposition 6.1] is proved for linearly many clauses, and the bound $2^{O(k^2 N \log\log N/\sqrt{\log N})}$ of [Cla26] for k-CNF is trivial once $k = \omega((\log N)^{1/4})$. The exponent 2 in Theorem 5.3 is a limit of the reduction, not of its analysis (Remark 5.4).

Remark 5.4 (The exponent 2). The construction of Lemma 5.2 applies to every one-message MA protocol for disjointness with perfect completeness and soundness 1/2, such as that of Imported Theorem 2.9: with p-bit proofs and ℓ-bit messages it gives 2^p instances of dimension at least 2^ℓ. Running such a protocol twice with independent coins and letting Alice announce her decision gives a protocol with $c_M = p$ and $c_P = 2\ell+1$ that computes DISJ_D with error 1/4, so $(p+1)(2\ell+1) \geq \alpha_0 D$ by Lemma D.1. Polynomially many instances of Lemma 5.2 of dimension $n^{o(1)}$ therefore force $D = o(\log^2 n)$, and Williams' instances with $D = \lceil (2\log n)^{c_0} \rceil$ reach polynomially many instances of dimension $n^{o(1)}$ through this construction only if $c_0 < 2$.

5.3 A barrier for the bounded-entry threshold

The reductions of Theorems 3.3, 3.6 and 5.3 and Proposition 5.1, and the reduction of [AS23], which composes Rubinstein's reduction from OV to approximate nearest neighbor with [AS23, Lemma 4.7], map each vector separately to at most 2^p query or key vectors, one for each value of an index that the reduction enumerates, and separate orthogonal from non-orthogonal pairs by a gap in the logits. We call such a map a gap embedding.

Definition 5.5 (Gap embedding)**.** Let $D, p \geq 0$ and $d \geq 1$ be integers and $\Delta > 0$. A (D,p,d,Δ)*-gap embedding* consists of maps $q, k : \{0,1\}^p \times \{0,1\}^D \to \mathbb{R}^d$ and $h : \{0,1\}^p \times \{0,1\}^D \to \mathbb{R}$ such that for all $a, b \in \{0,1\}^D$:

(G1) if $\langle a, b\rangle = 0$, then $\langle q(\mu,a), k(\mu,b)\rangle \geq h(\mu,a)$ for some μ;
(G2) if $\langle a, b\rangle \geq 1$, then $\langle q(\mu,a), k(\mu,b)\rangle \leq h(\mu,a) - \Delta$ for all μ.

Its *scale* is $\sigma = \max_{\mu,a} \|q(\mu,a)\|_2 \cdot \max_{\mu,b} \|k(\mu,b)\|_2$.

Reductions that make copies of the queries and of the keys independently are the special case in which the index is a pair $\mu = (\mu_1, \mu_2)$, q depends only on μ_1 and k only on μ_2. The scale bounds every logit, by the Cauchy–Schwarz inequality, and it is at most $d \cdot \max|q(\mu,a)_t| \cdot \max|k(\mu,b)_t|$; in the normalization $\exp(QK^\top/d)$ with entries in $[-B,B]$ it is at most B^2. Building attention instances from a gap embedding multiplies the number of queries and keys, or the number of attention calls, by at most 2^p, and in the product case by 2^{p_1} and 2^{p_2} with $p_1 + p_2 = p$; so subpolynomial blowup means $p = o(\log n)$ and polynomial blowup means $p = O(\log n)$. The reductions of Theorems 3.3 and 5.3 and Proposition 5.1 and of [HAPM+26], at constant precision, have $\Delta \geq \ln n - O(1)$, because the keys above the threshold must outweigh the others; those of Theorem 3.6 have $\Delta \geq \alpha \ln n$, and the reduction of [AS23, Lemma 4.7], at precision n^{-C_a}, has $e^\Delta > 4n$, so that the n far keys together stay below one near key. The counting reductions of Theorem 4.1 and [GHS+26, Theorem 5.7], read with negated queries, have $\Delta = \ln(1+1/r)$ and are not covered by the corollary below. The next theorem shows that a gap embedding with small scale is an efficient Merlin–Arthur protocol for disjointness, so Klauck's lower bound [Kla03] applies.

Theorem 5.6 (Barrier)**.** *There is an absolute constant $\alpha_2 > 0$ such that every (D, p, d, Δ)-gap embedding with $D \geq 1$ and scale σ satisfies $\Gamma = \sigma/\Delta \geq 1/2$ and*

$$\Gamma^2 \log(2+\Gamma) \;\geq\; \alpha_2 \frac{D}{p+1}.$$

In the protocol, Merlin names the copy μ, and the players estimate $\langle q(\mu, a), k(\mu, b)\rangle$ from $O(\Gamma^2)$ random-sign projections of $k(\mu, b)$, each sent with $O(\log(2+\Gamma))$ bits; the dimension d does not enter. Appendix D.3 gives the proof.

Corollary 5.7. *Let $c_0 > 1$, $\delta_0 > 0$ and $D = \lceil (2\log n)^{c_0} \rceil$. For all large n, every (D, p, d, Δ)-gap embedding with $p = O(\log n)$ and $\Delta \geq \delta_0 \ln n$ has scale $\sigma \geq \delta_0 (\log n)^{(1+c_0)/2}/O(\sqrt{\log\log n})$, where the O depends on c_0 and on the constant in $p = O(\log n)$. If every copy μ is written in the normalization of [AS23] with entries in $[-B, B]$, then $B \geq (\delta_0)^{1/2}(\log n)^{(1+c_0)/4}/O((\log\log n)^{1/4})$.*

Proof. Theorem 5.6 gives $\Gamma^2 \log(2+\Gamma) \geq \alpha_2 D/(p+1) = \Omega((\log n)^{c_0-1})$. If $\Gamma \geq (\log n)^{c_0}$ the claim holds; otherwise $\log(2+\Gamma) = O(\log\log n)$, so $\Gamma = \Omega((\log n)^{(c_0-1)/2}/\sqrt{\log\log n})$ and $\sigma = \Gamma\Delta \geq \delta_0 \Gamma \ln n$. For the entries, multiplying $q(\mu, \cdot)$ by $c_\mu > 0$ and $k(\mu, \cdot)$ by $1/c_\mu$ changes neither the logits nor (G1) and (G2), so the bound holds for the rescaled embedding; choosing c_μ to equalize $\max_a \|q(\mu, a)\|_2$ and $\max_b \|k(\mu, b)\|_2$ makes its scale equal to the largest product of these two maxima over μ, which is at most B^2 in that normalization. □

The spread and the entry bound come apart. Theorem 5.3 and Corollary 5.7 together say that after the refutation the two forms of the threshold have different bases. In terms of the spread, hardness at $(1+\theta)\ln n$ follows, in dimension $n^{o(1)}$, from CNF-SETH$_{c_0}$ for any $c_0 < 2$. In terms of the entry bound, gap embeddings with polynomial blowup and gap $\Delta = \Omega(\log n)$ cannot derive the threshold $B = \Theta(\sqrt{\log n})$ of [AS23] from OV in dimension $(\log n)^{c_0}$ for any $c_0 > 1$: they need $B \geq (\log n)^{(1+c_0)/4 - o(1)}$. The instances of Theorem 5.3 are consistent with this. They reach spread $(1+\theta)\ln n$ through vectors with up to $2^{O(s \log s)}$ nonzero coordinates, so their logits stay below $(1+\theta)\ln n$ while their scale, which Theorem 5.6 bounds from below, is not limited by the spread. At $D = O(\log n)$ nothing is lost: Remark B.1 attains scale $O(\log n)$. By Theorem 5.6, a gap embedding of OV in dimension D with $p = O(\log n)$, $\Delta = \Theta(\log n)$ and scale $O(\log n)$ exists only if $D = O(\log n)$. So for gap embeddings, the entry form of the threshold needs OV to be hard in dimension $O(\log n)$, which is Hypothesis 2.4. The barrier constrains the scale, not the spread, and it does not cover counting reductions with a small gap, adaptive reductions, or reductions from satisfiability that do not pass through OV.

6 Which attention lower bounds survive

Every SETH-based lower bound in Tables 2 and 3 reduces from OV, from k-OV, or from a gap problem built on them. Its basis after the refutation is determined by the dimension the reduction needs and by whether it needs a gap. Three cases cover every reduction from OV or k-OV in the tables.

(A) *Reductions that need* OV *in dimension* $\Theta(\log n)$. Reductions through Rubinstein's approximate nearest neighbor [Rub18], the distributed PCP of [ARW17], or Chen's dimensionality reduction [Che18] must enumerate Merlin's proofs or Chen's residues, and stay subquadratic only from OV in dimension $c\log n$; the counting reduction of [GHS+26, Theorem 5.7] needs precision $2^{-\Theta(d)}$; and some reductions, such as that of [SXY26, Theorem 4.7], start from OV in dimension $c\log n$ directly. They now rest on LDOV alone.

(B) *The same for k parties.* They rest on low-dimensional k-OV, which implies LDOV and which no open hypothesis about satisfiability is known to imply.

(C) *Exact embeddings.* Reductions from exact OV or k-OV that preserve the dimension up to a polynomial factor survive under CNF-SETH, in dimension $(\log n)^{O(1)}$ and with whatever precision the embedding needs there (Proposition 5.1).

Groups (D) and (E) of Table 3 collect the bounds based on other hypotheses. The statements in the second column are as published, under SETH unless noted; the precision is additive unless noted, and B is the entry bound in the normalization used by each paper.

Two observations. First, no result in the regime of the threshold, head dimension $\Theta(\log n)$ and logits of order $\log n$, survives under a hypothesis about satisfiability: all of group (A) needs OV in dimension $\Theta(\log n)$, through the gap machinery, through precision $2^{-\Theta(d)}$ for [GHS+26, Theorem 5.7], or by starting from OV in dimension $c \log n$, as [SXY26, Theorem 4.7] does, and every exact embedding in group (C) pays with polylogarithmic dimension, larger logits or superpolynomially small precision. Theorem 5.3 reaches the threshold spread $(1+\theta)\ln n$ at constant precision from satisfiability, at the price of clause exponent below 2 and dimension $n^{o(1)}$. Second, the refutation of the Sparse Max-SAT Hypotheses [Li26, Corollaries 4.4 and 4.5] removes the basis of the Max-SAT-based bounds for poly-attention and Strassen attention in [CPA26]; the SETH-based part of that paper moves to group (B). Together with the refutation of the hinted matrix–vector conjectures [AV26], these are the attention lower bounds whose hypotheses are now false (for [vSZ24], in the range $\tau < \tau_*$) rather than unsupported.

Result	Published statement	Route and status after the refutation
(A) OV in dimension $\Theta(\log n)$: LDOV only		
[AS23, Thm. 4.6]	attention, $d = C \log n$, $B = C_B\sqrt{\log n}$, precision n^{-C_a}	Rubinstein's gap ANN; LDOV. Constant precision and the exact constant in the spread: Theorem 3.3 and Corollary 3.7
[AS24a, Thm. 5.5], [HSK+25, Thm. 5.1]	gradient, and LoRA gradients, $d = O(\log n)$, $B = O(\sqrt{\log n}\, g(n))$ for any $g(n) = \omega(1)$ with $g(n) = o(\log n)$, precision $O(\log^{-4} n)$	reduce from the instance of [AS23, Lemma 4.7], the LoRA bound through [AS24a]; LDOV
[HAPM+26, Thms. 3.7, 4.1, 5.1]	every nontrivial constant precision; with KV-cache preprocessing; finding heavy keys; $d = C \log n$, $B = C\sqrt{\log n}$	Rubinstein's gap closest pair; LDOV
[GHS+26, Thm. 5.4]	$d = 2^{\Theta(\log^* n)}$, $B = \text{poly}(n)$, constant precision	Chen's dimensionality reduction, which needs $n^{o(1)}$ instances and hence OV in dimension $c \log n$; LDOV
[GHS+26, Thm. 5.7]	$d = C \log n$, $B = \ln 2$, precision n^{-C}	parity of $\sum_b 2^{\langle a,b\rangle}$; LDOV. In dimension $(\log n)^C$ the same argument needs precision $2^{-\Theta((\log n)^C)}$; Theorem 4.1 generalizes it
[AY25, Thms. 1.2, 1.3]	most similar documents: monochromatic with $\ell = n^{c/\log\log n}$, bichromatic with $\ell = c \log n$	additive approximate Max-IP [Che18]; LDOV
[AG24, Thms. 1.6, 1.8]	Gaussian and Student KDE in dimension $o(\log n)$	Chen's dimensionality reduction; LDOV
[AA22, Prop. 1.8]	batch Gaussian KDE, dimension $\alpha \log n$, error $n^{-\beta}$, squared diameter $O(\log n)$	approximate Hamming nearest neighbor [Rub18]; LDOV
[ACSS20, Cors. 5.28, 5.30, Prop. 5.31, Thms. 8.2, 8.6]	kernel matrix–vector products, sparsification and Laplacian solving in dimension $c^{\log^* n}$	Chen's dimensionality reduction; LDOV
[HLSL24, Thm. 2.1], [HWG+25, Thm. 3.1], [HWL+24, Prop. 4.1], [KLL+25, Thm. 4.5], [AS25, Thm. 1.4], [CHL+24, Thm. 6.1]	modern Hopfield models, prompt tuning, diffusion transformers, visual autoregressive models, RoPE attention and its training, at $d = O(\log n)$ and $B = \Theta(\sqrt{\log n})$ ($B = \omega(\sqrt{\log n})$ in [CHL+24])	special cases of [AS23] or [AS24a], or Rubinstein directly; LDOV
[SSWZ23, Thm. 3.3]	low-rank approximation of $f(UV)$	approximate Max-IP of [CW19]; LDOV
[SXY26, Thm. 4.7]	integer OV with $O(\log n)$-bit entries in every constructible dimension $D(n) = \omega(1)$, that is, $D(n)$ computable in time $n^{o(1)}$	from OV in dimension $c \log n$; LDOV
(B) Low-dimensional k-OV only		
[AS24b, Thm. B.2]	tensor attention, $n^{3-o(1)}$, $d = C \log n$, $B = \Theta((\log n)^{1/3})$	three-party gap Max-IP; low-dimensional 3-OV
[CLSS24, Thm. 5.3]	tensor attention gradients	from [AS24b]; low-dimensional 3-OV
[CPA26, Thm. E.3(1)]	poly-attention of degree k, $n^{k-o(1)}$, $B = \Omega((\log n)^{1/k})$	gap k-IP in dimension $c \log n$; low-dimensional k-OV

Table 2: SETH-based lower bounds for attention and related kernels that need low-dimensional OV or k-OV, and their basis after October 2026.

Result	Published statement	Route and status after the refutation
(C) Exact embeddings: CNF-SETH in dimension $(\log n)^{O(1)}$		
[KWH23, Section 4 and appendix]	self-attention, $d = \omega(\log n)$, exact, multiplicative or additive error	exact OV, dimension $\Theta(d)$; under CNF-SETH at $d = (\log n)^C$. The additive versions need precision $e^{-\Theta(d \log n + d^2)}$
[GHS+26, Lemmas 5.1, 5.6]	OV in dimension d reduces to attention with $B = O(\sqrt{\log n})$, constant precision	exact OV; under CNF-SETH at $d = (\log n)^C$ (Proposition 5.1)
[SXYY26, Thms. 3.1, 3.4]	with $L, H = \text{poly}(N)$, computing L layers of H heads to additive error $1/(10N)$ needs time $LHN^{2-o(1)}$ (Thm. 3.1); a transformer with linear-time embedding and MLP maps that computes 3-OV needs $LHm = N^{1-o(1)}$ (Thm. 3.4); $m = \Theta(\log N)$, under the 3-OV hypothesis	exact unbalanced 3-OV, $m = 2d + 2$; under CNF-SETH at $m = (\log N)^C$
[AY25, Thm. 1.1]	least similar documents, $\ell = c \log n$	exact OV; under CNF-SETH at $\ell = (\log n)^C$
[AG24, Thms. 3.2, 3.3]	KDE for a kernel f in dimension $m > C \log n$ at precision $(3n\tau(M))^{-1}$, where $\tau(M) = \|M^{-1}\|_{\infty\to\infty}$ for the $m \times m$ counting matrix M of f	exact closest pair; under CNF-SETH at $m = (\log n)^C$, at the same precision
[SSWZ23, Thms. 3.1, 3.2, 4.3]	low-rank approximation of $f(UV)$ (Thms. 3.1, 3.2) and computing $f(UV)z$ (Thm. 4.3)	exact OV; under CNF-SETH at $r = (\log n)^C$
[BIS17, Thms. 4, 6, 8]	kernel SVM, kernel ridge regression and kernel PCA with the Gaussian kernel, $d = \omega(\log n)$, multiplicative error $1 + \exp(-\omega(\log^2 n))$; additive error $\exp(-\omega(\log^2 n))$ in the proofs of the full version	exact bichromatic Hamming closest pair, to which OV reduces in the same dimension [AW15, Thm. 1.4]; under CNF-SETH at $d = (\log n)^C$, with additive error $e^{-\Omega(d \log n)}$
[ACSS20, Thms. 5.14, 5.15]	kernel matrix–vector products for $f(\|x - y\|^2)$ in dimension d on 1.01^d points, error $(\kappa(d+1))^{O(d^4)}$	exact bichromatic Hamming nearest neighbors (Lemma 5.25); under CNF-SETH at $d = (\log n)^C$
(D) Refuted bases		
[vSZ24, Lemmas 1.4, 5.5, 5.9]	dynamic attention, hinted matrix–vector conjecture	refuted by [AV26] for $\tau < \tau_*$, where $\tau_* = \ln 4/(5 \ln 10) = 0.12041\ldots$, and not refuted by it for $\tau \geq \tau_*$; [AV26, Section 5.4] discusses this bound
[CPA26, Thms. E.3(2),(3), C.10]	poly-attention and Strassen attention under the Sparse Max-kSAT and Max-2SAT hypotheses	both hypotheses are false [Li26, Corollaries 4.4 and 4.5]; no replacement known
(E) Not about satisfiability		
[GHS+26, Thm. 5.5]	$d = \text{poly}(n)$ needs $T_{\text{MUL}}(n, d, n)^{1-o(1)}$	generalized high-dimensional OV hypothesis; unaffected
[SXYY26, Thm. 5.1], [AY25, Thm. 1.4]	arithmetic-circuit lower bound; representational results	unconditional

Table 3: Lower bounds for attention and related kernels from exact embeddings and from other hypotheses, and their basis after October 2026.

7 Related work and open problems

Fine-grained complexity of attention. Keles, Wijewardena and Hegde [KWH23] gave the first SETH-based quadratic lower bounds for self-attention, for $d = \omega(\log n)$. Alman and Song [AS23] found the bounded-entry threshold and the polynomial-method algorithm, building on the Gaussian kernel density bounds of Aggarwal and Alman [AA22] and the linear algebra on geometric graphs of Alman, Chu, Schild and Song [ACSS20]. The threshold was extended to gradients [AS24a] and to tensor attention [AS24b]. Gupta, Huang, Saha, Xu and Ye [GHS+26] gave an $\tilde{O}(n^{2-1/d}\operatorname{polylog}(B/\varepsilon))$ algorithm for constant d and showed hardness for $d = 2^{\Theta(\log^* n)}$ with large entries, through Chen's dimensionality reduction [Che18]. Haverbeck et al. [HAPM+26] extended the bounded-entry lower bound to every nontrivial constant precision, to algorithms that preprocess the KV cache, and to finding the keys that carry most of the weight. Alman and Yu [AY25] proved lower bounds for subquadratic alternatives to attention, Chakrabarti, Pitassi and Alman [CPA26] for higher-order attention, and Saha, Xu, Ye and Yu [SXYY26] for multi-head and multi-layer transformers. Most of these lower bounds are stated under SETH; Section 6 records the hypothesis of each and its current basis. Sampling-based approximations, such as KDEformer [ZHDK23] and HyperAttention [HJK+24], come with guarantees in terms of column norms of the attention matrix and stable ranks; Proposition 3.1 is the simplest guarantee of this kind, and under LDOV, Corollary 3.7 shows that its exponent is optimal in terms of the spread. Hashing-based estimators beat uniform sampling for Gaussian kernel density estimation to relative precision ε. Let $\mu = \frac{1}{n}\sum_j e^{-a\|q-k_j\|^2}$ be the density at the query. Uniform sampling needs $O(\mu^{-1}\varepsilon^{-2})$ samples per query. The query time of [CS17] is $\mu^{-1/2}\varepsilon^{-2}$ times a factor that grows with the diameter of the data, and that of [CKNS20] is $\mu^{-0.173-o(1)}\varepsilon^{-2}$ for $\mu = n^{-\Theta(1)}$ and $\varepsilon \geq 1/\operatorname{polylog} n$. This density changes under transformations that leave attention unchanged: appending to every key a coordinate that brings all keys to a common norm R, and a zero coordinate to every query, keeps every logit, makes the softmax weights proportional to $e^{-\|q-k_j\|^2/2}$, and drives μ to 0 as R grows. So guarantees in terms of μ give no bound in terms of the spread, and in terms of the spread uniform sampling cannot be improved by a polynomial factor under LDOV.

Hardness of approximation in P. The gap instances behind Theorem 3.3 come from the distributed PCP framework of Abboud, Rubinstein and Williams [ARW17], made efficient by Rubinstein [Rub18] with algebraic-geometry codes; Chen [Che18] characterized the approximability of Max-IP and connected exact Max-IP to NP·UPP protocols, and Chen and Williams [CW19] showed that many approximate problems in dimension $O(\log n)$ are equivalent to OV. These works turn communication protocols into reductions. Theorem 5.6 goes the other way, from reductions to protocols, and uses the lower bound of Klauck [Kla03] to limit what gap embeddings can achieve. The non-reducibility results of Carmosino et al. [CGI+16] assume NSETH, which implies SETH and is therefore false. Theorem 5.6 is unconditional and covers every gap embedding, including the reduction of [AS23] and those of Theorems 3.3, 3.6 and 5.3 and Proposition 5.1.

After the refutation. The ledger of [Li26, Section 6] sorts the conditional lower bounds of learning theory by the hypothesis that survives; this paper carries out the attention row of that ledger. Alman and Vassilevska Williams [AV26] refuted the 3SUM and APSP hypotheses in the same month, so bounds based on them need new bases too. LDOV and CNF-SETH$_{c_0}$ have different support. No open hypothesis about satisfiability is known to imply LDOV; what supports it is the lack of algorithms, since the best algorithm for OV in dimension $c\log n$ takes time $n^{2-1/O(\log c)}$ [AWY15, CW16], and the equivalence of many approximate problems in dimension $O(\log n)$ with OV [CW19], which

makes LDOV the common basis of all of them. CNF-SETH$_{c_0}$ for $c_0 \in (1,2)$ is a statement about satisfiability, and the algorithms that refuted ETH do not reach it (Section 5.2).

Open problems.

Open Problem 1 (Dimension at polynomial precision)**.** In dimension $(\log n)^C$ with $C > 1$ and at precision $n^{-\beta}$ with $\beta > 1$, Corollary 4.2(b) gives hardness at spread $O(\log n/\log\log n)$ under LDOV, while Proposition 4.6 handles spread $(\log n)^{1-a}$ with $O(n^{1+\eta})$ operations only for $a \in (4(\beta+5)(C-1)/\eta, 1]$, and nothing for $C \geq 1+\eta/(4(\beta+5))$. Where is the threshold? And at precision $n^{-1+\delta}$, coarser than the $1/n$ that counting needs, does dimension still move it?

Open Problem 2 (The clause exponent)**.** Is CNF-SETH$_{c_0}$ plausible for some $c_0 \in (1,2)$? An algorithm for CNF-SAT with $N^{1+\delta}$ clauses in time $2^{(1-\Omega(1))N}$ for every $\delta > 0$ would refute the hypothesis of Theorem 5.3 for every $c_0 < 2$, and the threshold would again rest on LDOV alone; hardness would make it a consequence of satisfiability.

Open Problem 3 (Spread-efficient embeddings)**.** Theorem 5.6 bounds the scale of gap embeddings, not their spread. Is there a gap embedding of $\mathsf{OV}_{n,D}$ with $D = (2\log n)^2$, at most $n^{o(1)}$ copies and dimension $(\log n)^{O(1)}$, in which all logits of a query lie within $(1+o(1))\Delta$ of each other? Such an embedding would give hardness at spread $(1+o(1))\ln n$ from CNF-SETH$_2$ in polylogarithmic dimension. By Remark 5.4 it cannot come from the construction of Lemma 5.2. Without copies it would be a low-rank approximation of the disjointness matrix; with copies, a Merlin–Arthur analogue of approximate rank is needed.

Open Problem 4 (Lower-order terms and small dimension)**.** Under LDOV, at constant precision the exponent at spread $\kappa \ln n$ is $1+\min\{\kappa, 1\}$ (Corollary 3.7), but the lower bound loses a factor n^{γ} and needs dimension $C\log n$ with $C \to \infty$ as $\gamma \to 0$. Is the cost $n\,e^{W}(\log n)^{O(1)}$ for every $W \leq \ln n$? And what is the exponent in dimension $C\log n$ for a fixed C, where Theorem 3.3 gives hardness only at spread $(1+C^{-\Omega_\gamma(1)})\ln n$?

Open Problem 5 (Gradients)**.** Does the gradient of the attention loss have the same constant-precision threshold $\ln n$? The reduction of [AS24a] reads the attention output through the squared Frobenius norm of the whole output matrix, a sum over all rows, so the outputs of a NO instance must be $O(n^{-1/2})$, which raises the spread needed by the gadget of Theorem 3.3.

References

[AA22] Amol Aggarwal and Josh Alman. Optimal-degree polynomial approximations for exponentials and Gaussian kernel density estimation. In *37th Computational Complexity Conference (CCC)*, volume 234 of *LIPIcs*, pages 22:1–22:23, 2022. Proposition numbers refer to arXiv:2205.06249v1.

[ACSS20] Josh Alman, Timothy Chu, Aaron Schild, and Zhao Song. Algorithms and hardness for linear algebra on geometric graphs. In *61st IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 541–552, 2020. Theorem numbers refer to arXiv:2011.02466v1.

[AG24] Josh Alman and Yunfeng Guan. Finer-grained hardness of kernel density estimation. In *39th Computational Complexity Conference (CCC)*, volume 300 of *LIPIcs*, pages 35:1–35:21, 2024. Theorem numbers refer to arXiv:2407.02372v1.

[ARW17] Amir Abboud, Aviad Rubinstein, and Ryan Williams. Distributed PCP theorems for hardness of approximation in P. In *58th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 25–36, 2017.

[AS23] Josh Alman and Zhao Song. Fast attention requires bounded entries. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 63117–63135, 2023.

[AS24a] Josh Alman and Zhao Song. The fine-grained complexity of gradient computation for training large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, pages 61419–61454, 2024. Theorem numbers refer to arXiv:2402.04497v1.

[AS24b] Josh Alman and Zhao Song. How to capture higher-order correlations? Generalizing matrix softmax attention to Kronecker computation. In *International Conference on Learning Representations (ICLR)*, pages 4306–4333, 2024.

[AS25] Josh Alman and Zhao Song. Fast RoPE attention: Combining the polynomial method and fast Fourier transform. arXiv:2505.11892, 2025.

[AV26] Josh Alman and Virginia Vassilevska Williams. Truly subquadratic 3SUM and truly subcubic APSP via triangles in sparse lopsided graphs. arXiv:2610.06783, 2026.

[AW15] Josh Alman and Ryan Williams. Probabilistic polynomials and Hamming nearest neighbors. In *56th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 136–150, 2015. Theorem numbers refer to arXiv:1507.05106v1.

[AWY15] Amir Abboud, Ryan Williams, and Huacheng Yu. More applications of the polynomial method to algorithm design. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 218–230, 2015.

[AY25] Josh Alman and Hantao Yu. Fundamental limitations on subquadratic alternatives to transformers. In *International Conference on Learning Representations (ICLR)*, pages 47475–47498, 2025.

[BIS17] Arturs Backurs, Piotr Indyk, and Ludwig Schmidt. On the fine-grained complexity of empirical risk minimization: Kernel methods and neural networks. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 4308–4318, 2017.

[CGI+16] Marco L. Carmosino, Jiawei Gao, Russell Impagliazzo, Ivan Mihajlin, Ramamohan Paturi, and Stefan Schneider. Nondeterministic extensions of the strong exponential time hypothesis and consequences for non-reducibility. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS)*, pages 261–270, 2016.

[Che18] Lijie Chen. On the hardness of approximate and exact (bichromatic) maximum inner product. In *33rd Computational Complexity Conference (CCC)*, volume 102 of *LIPIcs*, pages 14:1–14:45, 2018. Lemma numbers refer to arXiv:1802.02325v2.

[CHL+24] Yang Cao, Jiayan Huo, Yingyu Liang, Zhenmei Shi, and Zhao Song. RoPE attention can be trained in almost linear time. arXiv:2412.17316, 2024.

[CKNS20] Moses Charikar, Michael Kapralov, Navid Nouri, and Paris Siminelakis. Kernel density estimation through density constrained near neighbor search. In *61st IEEE Symposium*

on Foundations of Computer Science (FOCS), pages 172–183, 2020. Theorem numbers refer to arXiv:2011.06997v1.

[Cla26] Claude (Anthropic). A $2^{o(n)}$ time algorithm for 3-SAT. Hexagon preprint hexagon:2610.00577v1, dated October 26, 2026. https://hexagonmath.org/2610.00577v1. Lean development at https://github.com/anthropics/formal-math/tree/main/eth, 2026.

[CLSS24] Yang Cao, Yingyu Liang, Zhenmei Shi, and Zhao Song. Training tensor attention efficiently: From cubic to almost linear time. arXiv:2405.16411, 2024.

[CPA26] Sayak Chakrabarti, Toniann Pitassi, and Josh Alman. Poly-attention: A general scheme for higher-order self-attention. In *International Conference on Learning Representations (ICLR)*, pages 11844–11906, 2026. arXiv:2602.02422.

[CS17] Moses Charikar and Paris Siminelakis. Hashing-based-estimators for kernel density in high dimensions. In *58th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 1032–1043, 2017. Theorem numbers refer to arXiv:1808.10530v1.

[CW16] Timothy M. Chan and Ryan Williams. Deterministic APSP, orthogonal vectors, and more: Quickly derandomizing Razborov–Smolensky. In *Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1246–1255, 2016.

[CW19] Lijie Chen and Ryan Williams. An equivalence class for orthogonal vectors. In *Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 21–40, 2019.

[GHS+26] Shreya Gupta, Boyang Huang, Barna Saha, Yinzhan Xu, and Christopher Ye. Subquadratic algorithms and hardness for attention with any temperature. In *International Conference on Learning Representations (ICLR)*, pages 85672–85701, 2026. Theorem numbers refer to arXiv:2505.14840v1.

[HAPM+26] Lukas Haverbeck, Carmen Amo Alonso, Andres Felipe Posada-Moreno, Sebastian Trimpe, and Marco Pavone. Efficiently approximating attention is hard. arXiv:2609.37261, 2026.

[HJK+24] Insu Han, Rajesh Jayaram, Amin Karbasi, Vahab Mirrokni, David P. Woodruff, and Amir Zandieh. HyperAttention: Long-context attention in near-linear time. In *International Conference on Learning Representations (ICLR)*, pages 39310–39325, 2024.

[HLSL24] Jerry Yao-Chieh Hu, Thomas Lin, Zhao Song, and Han Liu. On computational limits of modern Hopfield models: A fine-grained complexity analysis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pages 19327–19343, 2024. arXiv:2402.04520.

[HSK+25] Jerry Yao-Chieh Hu, Maojiang Su, En-Jui Kuo, Zhao Song, and Han Liu. Computational limits of low-rank adaptation (LoRA) fine-tuning for transformer models. In *International Conference on Learning Representations (ICLR)*, pages 6167–6206, 2025. Theorem numbers refer to arXiv:2406.03136v2.

[HWG+25] Jerry Yao-Chieh Hu, Wei-Po Wang, Ammar Gilani, Chenyang Li, Zhao Song, and Han Liu. Fundamental limits of prompt tuning transformers: Universality, capacity and efficiency. In *International Conference on Learning Representations (ICLR)*, pages 29634–29686, 2025. arXiv:2411.16525.

[HWL+24] Jerry Yao-Chieh Hu, Weimin Wu, Zhuoru Li, Sophia Pi, Zhao Song, and Han Liu. On statistical rates and provably efficient criteria of latent diffusion transformers (DiTs). In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, pages 31562–31628, 2024. arXiv:2407.01079.

[IPZ01] Russell Impagliazzo, Ramamohan Paturi, and Francis Zane. Which problems have strongly exponential complexity? *Journal of Computer and System Sciences*, 63(4):512–530, 2001.

[Kla03] Hartmut Klauck. Rectangle size bounds and threshold covers in communication complexity. In *18th Annual IEEE Conference on Computational Complexity (CCC)*, pages 118–134, 2003. Theorem numbers refer to arXiv:cs/0208006.

[KLL+25] Yekun Ke, Xiaoyu Li, Yingyu Liang, Zhizhou Sha, Zhenmei Shi, and Zhao Song. On computational limits and provably efficient criteria of visual autoregressive models: A fine-grained complexity analysis. arXiv:2501.04377, 2025.

[KWH23] Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. On the computational complexity of self-attention. In *International Conference on Algorithmic Learning Theory (ALT)*, volume 201 of *PMLR*, pages 597–619, 2023.

[Lam95] Leslie Lamport. How to write a proof. *The American Mathematical Monthly*, 102(7):600–608, 1995.

[Lam12] Leslie Lamport. How to write a 21st century proof. *Journal of Fixed Point Theory and Applications*, 11(1):43–63, 2012.

[Li26] Xiaoyu Li. What does the refutation of the Exponential Time Hypothesis mean for learning theory? Hexagon preprint hexagon:2610.00613v2, October 2026. https://hexagonmath.org/2610.00613v2, 2026.

[Ope26] OpenAI. The Exponential Time Hypothesis is false. OpenAI Math Release preprint, dated October 17, 2026, revised October 25, 2026, released October 26, 2026. https://github.com/openai/math/blob/main/preprints/The-Exponential-Time-Hypothesis-is-False-October-25-2026/paper.pdf, 2026.

[Raz92] Alexander A. Razborov. On the distributional complexity of disjointness. *Theoretical Computer Science*, 106(2):385–390, 1992.

[Rub18] Aviad Rubinstein. Hardness of approximate nearest neighbor search. In *Proceedings of the 50th Annual ACM Symposium on Theory of Computing (STOC)*, pages 1260–1268, 2018. Theorem numbers refer to arXiv:1803.00904v1.

[SSWZ23] Tamas Sarlos, Xingyou Song, David P. Woodruff, and Qiuyi Zhang. Hardness of low rank approximation of entrywise transformed matrix products. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 52568–52582, 2023. arXiv:2311.01960.

[SXY26] Barna Saha, Yinzhan Xu, and Christopher Ye. Furthest pair requires quadratic time in superconstant dimension under SETH. arXiv:2606.25887, 2026.

[SXYY26] Barna Saha, Yinzhan Xu, Christopher Ye, and Hantao Yu. On the computational hardness of transformers. In *Proceedings of the 58th Annual ACM Symposium on Theory of Computing (STOC)*, pages 1073–1084, 2026. Theorem numbers refer to arXiv:2603.11332v1.

[vSZ24] Jan van den Brand, Zhao Song, and Tianyi Zhou. Algorithm and hardness for dynamic attention maintenance in large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pages 49008–49028, 2024. Lemma numbers refer to arXiv:2304.02207v1.

[Wil05] Ryan Williams. A new algorithm for optimal 2-constraint satisfaction and its implications. *Theoretical Computer Science*, 348(2–3):357–365, 2005.

[ZHDK23] Amir Zandieh, Insu Han, Majid Daliri, and Amin Karbasi. KDEformer: Accelerating transformers via kernel density estimation. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pages 40605–40623, 2023.

A Roadmap of the appendix

The appendix proves the results of the main text, in the order of the sections, and its subsections follow those of the body. Eight short results are proved where they are stated: Corollaries 3.2, 3.7, 4.2, 4.5 and 5.7 and Remarks 3.5, 4.3 and 5.4. Figure 2 shows how the results depend on each other.

Appendix B starts with Lemma 2.3, which bounds how far a small change of the logits moves an output. Appendix B.1 analyzes uniform sampling with Bernstein's inequality (Proposition 3.1). Appendix B.2 builds attention instances with a dummy key from the gap instances of Imported Theorem 2.8 and proves hardness above the threshold (Theorem 3.3). Remark B.1 writes these instances in the normalization of [AS23], and Corollary 3.4 repeats the proof with the gap $\tau = \lceil \log n \rceil$, which reaches spread $\ln n + O(1)$ in polylogarithmic dimension. Appendix B.3 proves hardness below the threshold (Theorem 3.6): it applies the same gadget to a group of keys and copies every key, which leaves the outputs and the spread unchanged. Corollary 3.7 combines the two sides, and this gives Theorem A.

Appendix C.1 proves the precision lower bound (Theorem 4.1): attention at high precision reveals the number of orthogonal partners modulo $r+1$, where intersections are counted over blocks of coordinates. Appendix C.2 proves the polynomial method with the dependence on the precision and on the spread made explicit (Proposition 4.4), and Appendix C.3 carries that proof over to higher dimension (Proposition 4.6). Corollary 4.2 chooses the parameters, and with Corollary 4.5 this gives Theorem B.

Appendix D starts with the proof of Williams' reduction (Imported Theorem 2.7). Appendix D.1 combines it with an exact embedding (Proposition 5.1). Appendix D.2 turns Rubinstein's protocol (Imported Theorem 2.9) into gap instances (Lemma 5.2) and places them in the attention instance from the proof of Theorem 3.3. This proves Theorem 5.3 under CNF-SETH$_{c_0}$, without Hypothesis 2.4. These two results give Theorem C. Appendix D.3 derives Klauck's tradeoff between the proof length and the communication (Lemma D.1) from Razborov's corruption bound (Imported Theorem 2.10). Theorem 5.6 then turns a gap embedding into a Merlin–Arthur protocol for disjointness through random-sign projections, and with Corollary 5.7 this gives Theorem D. Remark 5.4 combines Lemma 5.2 with Lemma D.1 to show that the exponent 2 in Theorem 5.3 is a limit of the reduction.

How to read the proofs. We write the proofs in the structured style of Lamport [Lam95, Lam12]. A proof is a list of numbered steps, and $\langle k \rangle j$ is the j-th step at level k. Each step states a claim, and a proof of the claim follows it. This proof is a short paragraph or a list of steps at the next level. The last step of each list, Q.E.D., proves the goal of the list from the earlier steps. ASSUME and PROVE state the hypothesis and the goal of a step, and LET introduces a symbol. A reader who wants only the main idea can read the level-1 steps and skip their proofs. Each restated result keeps its number from the main text.

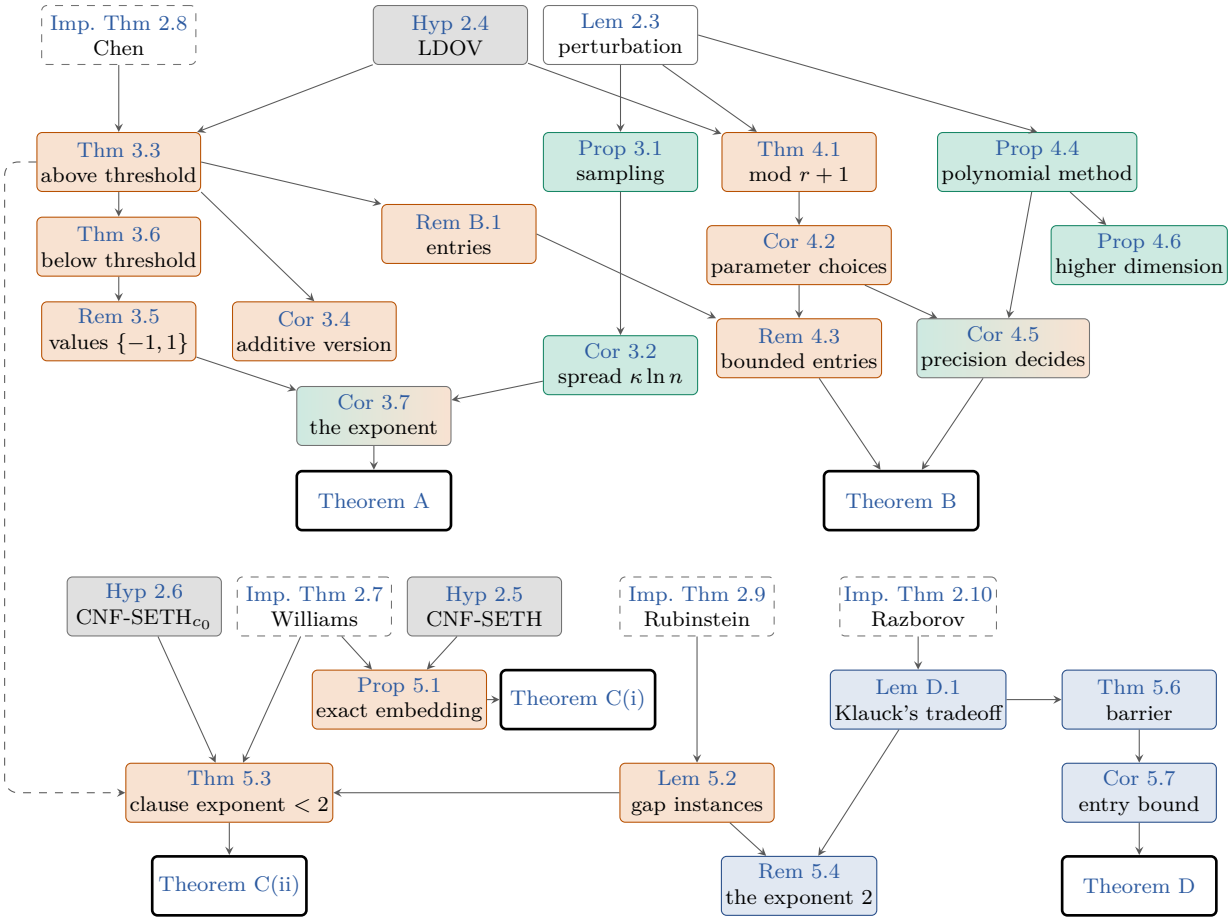


Figure 2: Dependencies between the results. An arrow from X to Y means that the proof of Y uses X; arrows implied by a longer path are omitted. The arrow into Theorem 5.3 is dashed because its proof reuses the attention instance from the proof of Theorem 3.3, not the statement, so it does not need Hypothesis 2.4. Gray boxes are hypotheses, dashed boxes imported theorems and thick boxes the theorems of the introduction. Green marks upper bounds, orange lower bounds and blue the barrier; a box with both green and orange states both bounds. Remark 3.5 also applies to Theorems 3.3 and 5.3 and Corollary 3.4, which gives Theorem C(ii) at every nontrivial constant precision; only its use in Corollary 3.7 is drawn.

B Proofs for Sections 2 and 3

Lemma 2.3 (Logit perturbation)**.** *Let two instances share their values and have logits ℓ_{ij} and ℓ'_{ij} with $|\ell_{ij} - \ell'_{ij}| \le \eta$ for all j and a fixed row i. Then $\|o_i - o'_i\|_\infty \le e^{2\eta} - 1$.*

Proof. ⟨1⟩1. For every j, $e^{-2\eta}\pi_{ij} \le \pi'_{ij} \le e^{2\eta}\pi_{ij}$.

PROOF: The numerator $e^{\ell'_{ij}}$ lies in $[e^{-\eta}e^{\ell_{ij}}, e^{\eta}e^{\ell_{ij}}]$, and so does the denominator $\sum_{j'} e^{\ell'_{ij'}}$ relative to $\sum_{j'} e^{\ell_{ij'}}$.

⟨1⟩2. $\sum_j |\pi'_{ij} - \pi_{ij}| \le e^{2\eta} - 1$.

PROOF: By ⟨1⟩1, $|\pi'_{ij} - \pi_{ij}| \le \max\{e^{2\eta} - 1, 1 - e^{-2\eta}\}\pi_{ij} = (e^{2\eta} - 1)\pi_{ij}$, and $\sum_j \pi_{ij} = 1$.

⟨1⟩3. Q.E.D.

PROOF: For every value coordinate c, $|o'_{ic} - o_{ic}| = |\sum_j (\pi'_{ij} - \pi_{ij})v_{jc}| \le \sum_j |\pi'_{ij} - \pi_{ij}|$, because

$|v_{jc}| \le 1$. Apply ⟨1⟩2.

□

B.1 Sampling

Proposition 3.1 (Sampling)**.** *Consider an attention instance with n queries, n' keys, dimension d and d_v value coordinates. Let $F \ge 1$ satisfy $F \ge n'\pi_{ij}$ for all i, j, and let $\varepsilon, \delta \in (0, 1]$. There is a randomized algorithm that, with probability at least $1 - \delta$, outputs $\tilde{o}_i$ with $\|\tilde{o}_i - o_i\|_\infty \le \varepsilon$ for all i, using $O(nm(d + d_v))$ operations with*

$$m = \left\lceil 40\, F \ln(2n(1 + d_v)/\delta)/\varepsilon^2 \right\rceil.$$

Every instance with spread at most W satisfies the hypothesis with $F = e^W$.

Proof. ⟨1⟩1. Let: $\eta = \varepsilon/4$. Fix a query i, Let: $M = \max_j \ell_{ij}$, $w_j = e^{\ell_{ij} - M} \in (0, 1]$, $\mu = \frac{1}{n'} \sum_j w_j$ and $\nu_c = \frac{1}{n'} \sum_j w_j v_{jc}$ for each value coordinate c. Then $o_{ic} = \nu_c/\mu$, $|\nu_c| \le \mu$ and $\mu \ge 1/F$.
Proof: $\pi_{ij} = w_j/(n'\mu)$, so $o_{ic} = \nu_c/\mu$, and $|\nu_c| \le \mu$ because $|v_{jc}| \le 1$. Some w_j equals 1, so $F \ge n' \max_j \pi_{ij} = 1/\mu$.

⟨1⟩2. Let: the algorithm draw, for each query i independently, indices $j_1, \ldots, j_m$ uniformly from $[n']$ with replacement, and output $\tilde{o}_{ic} = \sum_s e^{\ell_{ij_s}} v_{j_s c} / \sum_s e^{\ell_{ij_s}}$. It uses $O(nm(d + d_v))$ operations, and in exact arithmetic $\tilde{o}_{ic} = \hat{\nu}_c/\hat{\mu}$ with $\hat{\mu} = \frac{1}{m} \sum_s w_{j_s}$ and $\hat{\nu}_c = \frac{1}{m} \sum_s w_{j_s} v_{j_s c}$.
Proof: Each query computes m logits in $O(md)$ operations and two weighted sums in $O(md_v)$ operations. Multiplying numerator and denominator by e^{-M} does not change the ratio.

⟨1⟩3. For fixed i and c, each of the events $|\hat{\mu} - \mu| > \eta\mu$ and $|\hat{\nu}_c - \nu_c| > \eta\mu$ has probability at most $\delta/(n(1 + d_v))$.

⟨2⟩1. The summands $X_s = w_{j_s}$, respectively $X_s = w_{j_s} v_{j_s c}$, are independent, satisfy $|X_s| \le 1$, have mean μ, respectively ν_c, and variance at most μ.
Proof: $|X_s| \le w_{j_s} \le 1$, so $X_s^2 \le w_{j_s}^2 \le w_{j_s}$ and $\mathrm{Var}(X_s) \le \mathbb{E}[w_{j_s}] = \mu$.

⟨2⟩2. By Bernstein's inequality for independent variables with $|X_s - \mathbb{E}X_s| \le 2$, the probability that the empirical mean deviates by more than $t = \eta\mu$ is at most $2\exp(-mt^2/(2\mu + 4t/3)) \le 2\exp(-3m\eta^2\mu/7)$.
Proof: $mt^2/(2\mu + 4t/3) = m\eta^2\mu/(2 + 4\eta/3)$ and $2 + 4\eta/3 \le 7/3$ because $\eta \le 1/4$.

⟨2⟩3. Q.E.D.
Proof: By ⟨1⟩1, $\mu \ge 1/F$, so the bound of ⟨2⟩2 is at most $2\exp(-3m\varepsilon^2/(112F))$, and $m \ge 40F \ln(2n(1 + d_v)/\delta)/\varepsilon^2$ makes it at most $\delta/(n(1 + d_v))$, since $40 > 112/3$.

⟨1⟩4. With probability at least $1 - \delta$, $|\hat{\mu} - \mu| \le \eta\mu$ and $|\hat{\nu}_c - \nu_c| \le \eta\mu$ for all i and c.
Proof: A union bound over the $n(1 + d_v)$ events of ⟨1⟩3.

⟨1⟩5. On the event of ⟨1⟩4, $|\hat{\nu}_c/\hat{\mu} - \nu_c/\mu| \le 2\varepsilon/3$.
Proof: $|\hat{\nu}_c/\hat{\mu} - \nu_c/\mu| \le |\hat{\nu}_c - \nu_c|/\hat{\mu} + (|\nu_c|/\mu)\,|\mu - \hat{\mu}|/\hat{\mu} \le 2\eta\mu/((1 - \eta)\mu)$, using $|\nu_c| \le \mu$ and $\hat{\mu} \ge (1 - \eta)\mu$. With $\eta = \varepsilon/4 \le 1/4$ this is at most $(\varepsilon/2)(4/3) = 2\varepsilon/3$.

⟨1⟩6. Evaluating the exponentials to relative precision $\varepsilon/100$ changes $\tilde{o}_{ic}$ by at most $\varepsilon/3$.
Proof: Compute the m sampled logits to absolute error $\varepsilon/1000$ and subtract their maximum before exponentiating; then every weight lies in $(0, 1]$, and every weight is computed to relative error at most $e^{\varepsilon/1000}(1 + \varepsilon/100) - 1 \le \varepsilon/90$. The computed ratio is a weighted average of the same values with every weight multiplied by a factor in $[1 - \varepsilon/90, 1 + \varepsilon/90]$; as in Lemma 2.3, this moves the average by at most $(1 + \varepsilon/90)/(1 - \varepsilon/90) - 1 \le \varepsilon/40$. If $m \ge n'$, the algorithm

computes every output exactly instead, with $O(nn'(d+d_v)) \le O(nm(d+d_v))$ operations; so assume $m < n'$. The remaining m-term sums and the division are carried out to relative precision $\varepsilon/(1000m)$ per operation, on numbers of $O(\log(nm/\varepsilon)) = O(\log(nn'/\varepsilon))$ bits. Every partial sum of the numerator is at most $\sum_s w_s$ in absolute value, because $|v_{j_s c}| \le 1$, so the numerator and the denominator are each computed within $(\varepsilon/1000)\sum_s w_s$, and the ratio within $3\varepsilon/1000 + \varepsilon/(1000m) \le \varepsilon/100$.

⟨1⟩7. An instance with spread at most W satisfies $n'\pi_{ij} \le e^W$.
PROOF: Every $w_j \ge e^{-W}$, so $\mu \ge e^{-W}$ and $n'\pi_{ij} = w_j/\mu \le e^W$.

⟨1⟩8. Q.E.D.
PROOF: By ⟨1⟩4, ⟨1⟩5 and ⟨1⟩6, with probability at least $1-\delta$ every output is within $2\varepsilon/3+\varepsilon/3 = \varepsilon$; ⟨1⟩2 counts the operations and ⟨1⟩7 gives the last sentence.

□

B.2 Quadratic time above the threshold

Theorem 3.3 (Hardness above the threshold)**.** *Assume Hypothesis 2.4. For all $\gamma \in (0,1)$ and $\theta \in (0,1]$ there is a constant $C \ge 1$ such that for every $\varepsilon \in (0,1/2)$, $\mathsf{ATT}(n, \lceil C\log n\rceil, (1+\theta)\ln n, \varepsilon)$ cannot be solved in time $O(n^{2-\gamma})$. This holds already for instances with $d_v = 1$, values in $\{0,1\}$, and all logits in $[0,(1+\theta)\ln n]$.*

Proof. ⟨1⟩1. LET: $c \ge 1$ be the integer of Hypothesis 2.4 for $\gamma/2$, LET: $\tau = \lceil 4/\theta\rceil$, and apply Imported Theorem 2.8 with parameters c, $\gamma/4$ and τ; it gives n_0, a dimension $d_C \le \tau^{\mathrm{poly}(4c/\gamma)}\log n$ and an integer $T \ge c\log n$. LET: $C = 2\tau^{\mathrm{poly}(4c/\gamma)}$, so that $d_C + 1 \le \lceil C\log n\rceil$, and LET: $\eta = (1-2\varepsilon)/4$.
PROOF: $1 \le \tau^{\mathrm{poly}(4c/\gamma)}\log n$ for $n \ge 2$. Also $\eta \in (0, 1/4]$.

⟨1⟩2. ASSUME: an algorithm $\mathcal{A}$ solves $\mathsf{ATT}(n, \lceil C\log n\rceil, (1+\theta)\ln n, \varepsilon)$ in time $O(n^{2-\gamma})$ on the instances described in the theorem. PROVE: $\mathsf{OV}_{n,\lceil c\log n\rceil}$ can be solved in time $O(n^{2-\gamma/2})$, which contradicts the choice of c.

⟨1⟩3. LET: (A_ψ, B_ψ) be one of the instances of Imported Theorem 2.8, LET: ξ be a rational with $\ln((1-\eta)/\eta) \le \xi \le \ln((1-\eta)/\eta) + 1/10$, LET: λ be a rational of $O(\log n)$ bits with $\ln n + 2\ln(1/\eta) + 1/10 \le \lambda T(1-1/\tau) \le \ln n + 2\ln(1/\eta) + 1/5$, computed from rational approximations of $\ln n$ and $\ln(1/\eta)$, and LET: $t^* = \lambda T - \xi$. LET: the attention instance I_ψ have the queries $q_x = (\lambda x, t^*)$ for $x \in A_\psi$ and one zero query, the keys $k_y = (y, 0)$ with value 1 for $y \in B_\psi$, and the dummy key $k^* = (0,1)$ with value 0.
PROOF: I_ψ has $n+1$ queries, $n+1$ keys, dimension $d_C + 1$, $d_v = 1$, values in $\{0,1\}$ and rational entries of $O(\log n)$ bits.

⟨1⟩4. Every logit of I_ψ lies in $[0, \lambda T]$, and $\lambda T \le (1+\theta)\ln(n+1)$ for all $n \ge n_1(\theta,\varepsilon)$.

⟨2⟩1. $\ln n + 2\ln(1/\eta) + 1/10 \le \lambda T(1-1/\tau)$ and $\lambda T \le (\ln n + 2\ln(1/\eta) + 1/5)/(1-1/\tau)$.
PROOF: By the choice of λ.

⟨2⟩2. The logits of q_x are $\lambda\langle x, y\rangle \in [0, \lambda T]$ and $t^* \in [0, \lambda T]$; the zero query has all logits 0.
PROOF: Imported Theorem 2.8(i) gives $\langle x, y\rangle \in [0, T]$. By ⟨2⟩1, $\lambda T \ge \ln(1/\eta) + 1/10 \ge \xi$, since $\xi \le \ln((1-\eta)/\eta) + 1/10$; so $t^* \ge 0$.

⟨2⟩3. $\lambda T \le (1+\theta/2)(\ln n + 2\ln(1/\eta) + 1/5)$.
PROOF: By ⟨2⟩1 and $1/(1-1/\tau) \le 1 + 2/\tau \le 1 + \theta/2$, because $1/\tau \le \theta/4 \le 1/4$.

⟨2⟩4. Q.E.D.
PROOF: By ⟨2⟩2 and ⟨2⟩3, since $(1+\theta/2)(2\ln(1/\eta) + 1/5) \le (\theta/2)\ln n$ for $n \ge n_1(\theta,\varepsilon)$.

⟨1⟩5. LET: o_x be the output of q_x in I_ψ. If some $y \in B_\psi$ has $\langle x, y\rangle = T$, then $o_x \geq 1 - \eta$. If every $y \in B_\psi$ has $\langle x, y\rangle \leq T/\tau$, then $o_x \leq \eta$.

⟨2⟩1. $o_x = S_x/(S_x + e^{t^*})$ with $S_x = \sum_{y \in B_\psi} e^{\lambda\langle x,y\rangle}$.

PROOF: The real keys have value 1 and the dummy key has value 0.

⟨2⟩2. If $\langle x, y\rangle = T$ for some y, then $o_x \geq 1 - \eta$.

PROOF: $S_x \geq e^{\lambda T} = e^\xi e^{t^*}$, so $o_x \geq e^\xi/(e^\xi + 1) \geq 1 - \eta$, as $e^\xi \geq (1 - \eta)/\eta$.

⟨2⟩3. If $\langle x, y\rangle \leq T/\tau$ for all y, then $o_x \leq \eta$.

PROOF: $o_x \leq S_x/e^{t^*} \leq n \exp(\lambda T/\tau - \lambda T + \xi) = ne^\xi e^{-\lambda T(1-1/\tau)}$. By ⟨1⟩4, this is at most $e^{\xi - 1/10}\eta^2 \leq ((1 - \eta)/\eta)\eta^2 \leq \eta$.

⟨2⟩4. Q.E.D.

PROOF: By ⟨2⟩1, ⟨2⟩2 and ⟨2⟩3.

⟨1⟩6. LET: $N_\psi \leq 2n^{\gamma/4}$ be the number of instances. LET: the algorithm for OV solve inputs with $n < \max\{n_0, n_1, \tau\}$ by brute force; otherwise it computes the instances of Imported Theorem 2.8, builds each I_ψ, runs $\mathcal{A}$ on it $R = \lceil 18 \ln(3N_\psi)\rceil$ times, and declares ψ positive if in a majority of the runs some query q_x receives an output above $1/2$. It answers YES if and only if some ψ is positive. It errs with probability at most $1/3$.

⟨2⟩1. A run of $\mathcal{A}$ that succeeds reports an output above $1/2$ for every query with $o_x \geq 1 - \eta$, and for no query with $o_x \leq \eta$. The zero query is ignored.

PROOF: A successful run is within ε of every output, and $1 - \eta - \varepsilon > 1/2 > \eta + \varepsilon$, because $\eta + \varepsilon = (1 + 2\varepsilon)/4 < 1/2$.

⟨2⟩2. If (A, B) has an orthogonal pair, some instance ψ^* has a query with $o_x \geq 1 - \eta$. Otherwise every query of every I_ψ has $o_x \leq \eta$.

PROOF: Imported Theorem 2.8(ii) and (iii) with ⟨1⟩5.

⟨2⟩3. If (A, B) has an orthogonal pair, ψ^* is declared positive with probability at least $1 - 1/(3N_\psi)$. If not, each ψ is declared positive with probability at most $1/(3N_\psi)$.

PROOF: Each run succeeds independently with probability at least $2/3$. By Hoeffding's inequality, at least half of the R runs fail with probability at most $e^{-R/18} \leq 1/(3N_\psi)$. If a majority succeeds, the declaration about ψ^*, respectively about ψ, is the one given by ⟨2⟩1 and ⟨2⟩2.

⟨2⟩4. Q.E.D.

PROOF: On YES inputs, by ⟨2⟩3 for ψ^*. On NO inputs, by ⟨2⟩3 and a union bound over the N_ψ instances.

⟨1⟩7. The algorithm of ⟨1⟩6 runs in time $O(n^{2-\gamma/2})$.

PROOF: Imported Theorem 2.8 takes time $n^{1+\gamma/4}\text{poly}(\log n)$, building the instances takes $N_\psi O(nd_{\text{C}})$, and the calls take $N_\psi R \cdot O((n + 1)^{2-\gamma}) = O(n^{2-3\gamma/4} \log n)$. By ⟨1⟩1 and ⟨1⟩4, each I_ψ is an instance of $\mathsf{ATT}(n + 1, \lceil C \log(n + 1)\rceil, (1 + \theta)\ln(n + 1), \varepsilon)$ of the stated form, so the assumption on $\mathcal{A}$ applies.

⟨1⟩8. Q.E.D.

PROOF: ⟨1⟩6 and ⟨1⟩7 contradict Hypothesis 2.4 for $\gamma/2$, so the algorithm $\mathcal{A}$ of ⟨1⟩2 does not exist.

□

Remark B.1. The instances I_ψ in the proof of Theorem 3.3 can be written in the normalization $\exp(QK^\top/d')$ of [AS23] with all entries in $[-B, B]$ for $B^2 = O_{\gamma,\theta,\varepsilon}(\ln n)$ and $d' = O_{\gamma,\theta}(\log n)$.

Proof. ⟨1⟩1. Replace the last coordinate by a block of $T+1$ coordinates. The query q_x gets λu there, where $u \in [0,1]^{T+1}$ has coordinate sum $t^*/\lambda \le T$; the dummy key gets the all-ones vector; the real keys get zeros. All logits are unchanged.

PROOF: $\langle \lambda u, \mathbf{1} \rangle = t^*$, and the real keys contribute nothing in the new block.

⟨1⟩2. Now every query entry lies in $[0,\lambda]$ and every key entry in $\{0,1\}$. LET: $d' = d_{\mathrm{C}} + T + 1$, LET: s be a rational with $\sqrt{d'/\lambda} \le s \le 2\sqrt{d'/\lambda}$, and LET: $Q' = sQ$ and $K' = (d'/s)K$. Then $Q'K'^{\top}/d' = QK^{\top}$, and all entries of Q' and K' lie in $[-B,B]$ for $B^2 = 4d'\lambda$.

PROOF: The entries of Q' are at most $s\lambda \le 2\sqrt{d'\lambda}$, and those of K' at most $d'/s \le \sqrt{d'\lambda}$.

⟨1⟩3. Q.E.D.

PROOF: $d'\lambda \le (d_{\mathrm{C}} + T + 1)(\ln n + 2\ln(1/\eta) + 1/5)/(T(1-1/\tau))$ by step ⟨2⟩1 of ⟨1⟩4 in the proof of Theorem 3.3, and $(d_{\mathrm{C}} + T + 1)/T \le \tau^{\mathrm{poly}(4c/\gamma)}/c + 2$ depends only on γ and θ, because $T \ge c \log n$. Finally $d' \le 2d_{\mathrm{C}} + 1 = O_{\gamma,\theta}(\log n)$, because $T \le d_{\mathrm{C}}$.

□

Corollary 3.4 (Additive version)**.** *Assume Hypothesis 2.4. For all $\gamma \in (0,1)$ and $\varepsilon \in (0,1/2)$ there is $C \ge 1$ such that $\mathsf{ATT}(n, \lceil (\log n)^C \rceil, \ln n + 2\ln\frac{4}{1-2\varepsilon} + 2, \varepsilon)$ cannot be solved in time $O(n^{2-\gamma})$. For $\varepsilon = 1/6$ the spread is below $\ln n + 6$. This holds already for instances with $d_v = 1$ and values in $\{0,1\}$.*

Proof. ⟨1⟩1. Repeat the proof of Theorem 3.3 with $\tau = \lceil \log n \rceil$ in ⟨1⟩1, for $n \ge 16$. Then $d_{\mathrm{C}} \le \tau^{\mathrm{poly}(4c/\gamma)} \log n \le (\log n)^{C-1}$ for a constant C and large n.

PROOF: Imported Theorem 2.8 holds for every real $\tau \ge 4$ once $n \ge n_0$, where n_0 does not depend on τ, and its dimension bound is explicit in τ.

⟨1⟩2. All logits lie in $[0, \lambda T]$ with $\lambda T \le \ln n + 2\ln(1/\eta) + 2$ for large n.

PROOF: By step ⟨2⟩1 of ⟨1⟩4 in that proof, $\lambda T \le (\ln n + 2\ln(1/\eta) + 1/5)(1 + 2/\log n)$, which is $\ln n + 2\ln(1/\eta) + 1/5 + 2\ln 2 + o(1)$. Finally $2\ln(1/\eta) = 2\ln(4/(1-2\varepsilon))$, and for $\varepsilon = 1/6$ the bound is $\ln n + 2\ln 6 + 2 < \ln n + 6$.

⟨1⟩3. Q.E.D.

PROOF: Steps ⟨1⟩5 and ⟨1⟩6 of that proof use only items (i)–(iii) of Imported Theorem 2.8 and the bound $\lambda T(1-1/\tau) \ge \ln n + 2\ln(1/\eta) + 1/10$, so they hold unchanged; the condition $1/\tau \le \theta/4$ was used only for the spread bound, which ⟨1⟩2 replaces. In step ⟨1⟩7, Imported Theorem 2.8 now takes time $n^{1+\gamma/4}\mathrm{poly}(\log n)$ because d_{C} is polylogarithmic, building the instances takes $N_\psi O(nd_{\mathrm{C}}) = O(n^{1+\gamma/4}\mathrm{poly}(\log n))$, and the calls take $O(n^{2-3\gamma/4}\log n)$ as before; by ⟨1⟩1 and ⟨1⟩2, each I_ψ is an instance of $\mathsf{ATT}(n+1, \lceil (\log(n+1))^C \rceil, \ln(n+1) + 2\ln(1/\eta) + 2, \varepsilon)$ of the stated form. Step ⟨1⟩8 applies verbatim.

□

B.3 The exponent below the threshold

Theorem 3.6 (Hardness below the threshold)**.** *Assume Hypothesis 2.4. For all $\alpha \in (0,1]$, $\gamma \in (0,\alpha)$ and $\theta \in (0,1]$ there is a constant $C \ge 1$ such that for every $\varepsilon \in (0,1/2)$, $\mathsf{ATT}(n, \lceil C \log n \rceil, (1+\theta)\alpha \ln n, \varepsilon)$ cannot be solved in time $O(n^{1+\alpha-\gamma})$. This holds already for instances with $d_v = 1$, values in $\{0,1\}$, and all logits in $[0, (1+\theta)\alpha \ln n]$.*

Proof. We may assume that α is rational. For real α, choose a rational $\tilde{\alpha} \in (\alpha - \gamma, \alpha]$ and let $\tilde{\gamma} = \gamma - \alpha + \tilde{\alpha}$, so that $0 < \tilde{\gamma} < \tilde{\alpha}$; the rational case gives C such that spread $(1+\theta)\tilde{\alpha} \ln n \le (1+\theta)\alpha \ln n$ has no $O(n^{1+\tilde{\alpha}-\tilde{\gamma}}) = O(n^{1+\alpha-\gamma})$ algorithm in dimension $\lceil C \log n \rceil$, which is the claim for α.

⟨1⟩1. LET: c, $\tau = \lceil 4/\theta \rceil$, n_0, d_{C}, T, C and η be as in step ⟨1⟩1 of the proof of Theorem 3.3, so that c is the integer of Hypothesis 2.4 for $\gamma/2$, Imported Theorem 2.8 is applied with parameters c, $\gamma/4$ and τ, and $d_{\mathrm{C}} + 1 \le \lceil C \log n \rceil$. LET: $x(\psi, \cdot)$ and y be the maps of Imported Theorem 2.8(iv).
PROOF: As there.

⟨1⟩2. ASSUME: an algorithm $\mathcal{A}$ solves $\mathsf{ATT}(N, \lceil C \log N \rceil, (1+\theta)\alpha \ln N, \varepsilon)$ in time $O(N^{1+\alpha-\gamma})$ on the instances described in the theorem. PROVE: $\mathsf{OV}_{n,\lceil c \log n \rceil}$ can be solved in time $O(n^{2-\gamma/2})$, which contradicts the choice of c.

⟨1⟩3. LET: (A, B) be an instance of $\mathsf{OV}_{n,\lceil c \log n \rceil}$ and $n_B = \lceil n^\alpha \rceil$. Partition B into J groups $B_1, \ldots, B_J$ of at most n_B vectors each, and for each j LET: $M_j = \lceil n/(|B_j|+1) \rceil$ and $N_j = M_j(|B_j|+1)$. Then $J \le 2n^{1-\alpha}$ and $n \le N_j \le 2n$.
PROOF: $J = \lceil n/n_B \rceil \le n^{1-\alpha} + 1$. $N_j \ge n$ by the choice of M_j, and $N_j < n + |B_j| + 1$, so the integer N_j is at most $n + |B_j| \le 2n$.

⟨1⟩4. LET: ξ be a rational with $\ln((1-\eta)/\eta) \le \xi \le \ln((1-\eta)/\eta) + 1/10$, LET: λ be a rational of $O(\log n)$ bits with $\alpha \ln n + 2\ln(1/\eta) + 1 \le \lambda T(1-1/\tau) \le \alpha \ln n + 2\ln(1/\eta) + 11/10$, computed from rational approximations of $\ln n$ and $\ln(1/\eta)$, and LET: $t^* = \lambda T - \xi$. For each j and ψ LET: the instance $I_{j,\psi}$ have the queries $q_a = (\lambda x(\psi, a), t^*)$ for $a \in A$ and $N_j - n$ zero queries; M_j keys equal to $(y(b), 0)$ with value 1 for each $b \in B_j$; and M_j keys equal to $(0, 1)$ with value 0.
PROOF: $I_{j,\psi}$ has N_j queries, N_j keys, dimension $d_{\mathrm{C}} + 1$, $d_v = 1$, values in $\{0, 1\}$ and rational entries of $O(\log n)$ bits, and it is built in time $O(N_j d_{\mathrm{C}})$ from the vectors $x(\psi, a)$ and $y(b)$.

⟨1⟩5. Every logit of $I_{j,\psi}$ lies in $[0, \lambda T]$, and $\lambda T \le (1+\theta)\alpha \ln n \le (1+\theta)\alpha \ln N_j$ for all $n \ge n_1(\alpha, \theta, \varepsilon)$.

⟨2⟩1. $\alpha \ln n + 2\ln(1/\eta) + 1 \le \lambda T(1-1/\tau) \le \alpha \ln n + 2\ln(1/\eta) + 11/10$.
PROOF: By the choice of λ.

⟨2⟩2. The logits of q_a are $\lambda \langle x(\psi, a), y(b) \rangle \in [0, \lambda T]$ and $t^* \in [0, \lambda T]$; zero queries have all logits 0.
PROOF: Imported Theorem 2.8(i) and (iv) give $\langle x(\psi, a), y(b) \rangle \in [0, T]$. By ⟨2⟩1, $\lambda T \ge \ln(1/\eta) + 1 \ge \xi$, since $\xi \le \ln((1-\eta)/\eta) + 1/10$; so $t^* \ge 0$.

⟨2⟩3. $\lambda T \le (1 + \theta/2)(\alpha \ln n + 2\ln(1/\eta) + 11/10)$.
PROOF: By ⟨2⟩1 and $1/(1-1/\tau) \le 1 + 2/\tau \le 1 + \theta/2$, because $1/\tau \le \theta/4 \le 1/4$.

⟨2⟩4. Q.E.D.
PROOF: By ⟨2⟩2 and ⟨2⟩3, since $(1+\theta/2)(2\ln(1/\eta)+11/10) \le (\theta/2)\alpha \ln n$ for $n \ge n_1(\alpha, \theta, \varepsilon)$, and $N_j \ge n$.

⟨1⟩6. LET: o_a be the output of q_a in $I_{j,\psi}$. If some $b \in B_j$ has $\langle x(\psi, a), y(b) \rangle = T$, then $o_a \ge 1 - \eta$. If every $b \in B_j$ has $\langle x(\psi, a), y(b) \rangle \le T/\tau$, then $o_a \le \eta$.

⟨2⟩1. $o_a = S_a/(S_a + e^{t^*})$ with $S_a = \sum_{b \in B_j} e^{\lambda \langle x(\psi, a), y(b) \rangle}$.
PROOF: Every key occurs M_j times, so every term of the numerator and of the denominator of the softmax is multiplied by M_j. The real keys have value 1 and the dummy keys value 0.

⟨2⟩2. If $\langle x(\psi, a), y(b) \rangle = T$ for some $b \in B_j$, then $o_a \ge 1 - \eta$.
PROOF: $S_a \ge e^{\lambda T} = e^{\xi} e^{t^*}$, so $o_a \ge e^{\xi}/(e^{\xi} + 1) \ge 1 - \eta$, as $e^{\xi} \ge (1-\eta)/\eta$.

⟨2⟩3. If $\langle x(\psi, a), y(b) \rangle \le T/\tau$ for all $b \in B_j$, then $o_a \le \eta$.
PROOF: $o_a \le S_a/e^{t^*} \le n_B \exp(\lambda T/\tau - \lambda T + \xi) = n_B e^{\xi} e^{-\lambda T(1-1/\tau)}$. By ⟨2⟩1 of ⟨1⟩5 and $n_B \le 2n^\alpha$, this is at most $2e^{\xi-1}\eta^2 \le 2e^{-9/10}((1-\eta)/\eta)\eta^2 \le \eta$, since $2e^{-9/10} < 1$.

⟨2⟩4. Q.E.D.
PROOF: By ⟨2⟩1, ⟨2⟩2 and ⟨2⟩3.

⟨1⟩7. LET: $N_\psi \le 2n^{\gamma/4}$ be the number of indices ψ. LET: the algorithm for OV solve inputs with $n < \max\{n_0, n_1, \tau\}$ by brute force; otherwise it computes $x(\psi, a)$ and $y(b)$ for all ψ, $a \in A$ and $b \in B$, builds every $I_{j,\psi}$, runs $\mathcal{A}$ on it $R = \lceil 18 \ln(3JN_\psi) \rceil$ times, and declares (j, ψ) positive if in a majority of the runs some query q_a receives an output above $1/2$. It answers YES if and only if some (j, ψ) is positive. It errs with probability at most $1/3$.

⟨2⟩1. A run of $\mathcal{A}$ that succeeds reports an output above $1/2$ for every query with $o_a \ge 1 - \eta$, and for no query with $o_a \le \eta$. The zero queries are ignored.
PROOF: $1 - \eta - \varepsilon > 1/2 > \eta + \varepsilon$, because $\eta + \varepsilon = (1 + 2\varepsilon)/4 < 1/2$.

⟨2⟩2. If $\langle a^*, b^* \rangle = 0$ with $a^* \in A$ and $b^* \in B_{j^*}$, then some I_{j^*,ψ^*} has $o_{a^*} \ge 1 - \eta$. If (A, B) has no orthogonal pair, every query of every $I_{j,\psi}$ has $o_a \le \eta$.
PROOF: Imported Theorem 2.8(iv) and ⟨1⟩6.

⟨2⟩3. In the first case (j^*, ψ^*) is declared positive with probability at least $1 - 1/(3JN_\psi)$. In the second, each (j, ψ) is declared positive with probability at most $1/(3JN_\psi)$.
PROOF: Each run succeeds independently with probability at least $2/3$. By Hoeffding's inequality, at least half of the R runs fail with probability at most $e^{-R/18} \le 1/(3JN_\psi)$. If a majority succeeds, the declaration is the one given by ⟨2⟩1 and ⟨2⟩2.

⟨2⟩4. Q.E.D.
PROOF: On YES inputs, by ⟨2⟩3 for (j^*, ψ^*). On NO inputs, by ⟨2⟩3 and a union bound over the JN_ψ instances.

⟨1⟩8. The algorithm of ⟨1⟩7 runs in time $O(n^{2-\gamma/2})$.
PROOF: Imported Theorem 2.8 takes time $n^{1+\gamma/4}\mathrm{poly}(\log n)$. Building the instances takes $JN_\psi \cdot O(nd_{\mathrm{C}}) = O(n^{2-\alpha+\gamma/4} \log n)$, and the calls take $JN_\psi R \cdot O((2n)^{1+\alpha-\gamma}) = O(n^{2-3\gamma/4} \log n)$. Both are $O(n^{2-\gamma/2})$, the first because $\gamma < \alpha$. By ⟨1⟩1 and ⟨1⟩5, after padding with zero coordinates each $I_{j,\psi}$ is an instance of $\mathsf{ATT}(N_j, \lceil C \log N_j \rceil, (1+\theta)\alpha \ln N_j, \varepsilon)$ of the stated form, so the assumption on $\mathcal{A}$ applies.

⟨1⟩9. Q.E.D.
PROOF: ⟨1⟩7 and ⟨1⟩8 contradict Hypothesis 2.4 for $\gamma/2$, so the algorithm $\mathcal{A}$ of ⟨1⟩2 does not exist.

□

C Proofs for Section 4

C.1 Counting modulo $r + 1$

Theorem 4.1 (Precision lower bound)**.** *Assume Hypothesis 2.4. For every $\gamma \in (0, 1)$ there is an integer $c \ge 1$ with the following property. Let $r, w : \mathbb{N} \to \mathbb{N}$ be computable in time $O(n)$, with $r(n), w(n) \ge 1$ for all n, and $r(n) \le n$ and $w(n) \le \frac{1}{2} \log n$ for all large n. Let*

$$K = \lceil \lceil c \log n \rceil / w \rceil, \qquad d = K(2^w - 1), \qquad W = K \ln(1 + 1/r), \qquad \varepsilon = \frac{1}{64\, n\, (r+1)^{2K}},$$

with $r = r(n)$ and $w = w(n)$. Then $\mathsf{ATT}(n, d, W, \varepsilon)$ cannot be solved in time $O(n^{2-\gamma})$. This holds already for $d_v = 1$, values in $\{0, 1\}$, and all logits in $[0, W]$.

Proof. ⟨1⟩1. LET: $c \ge 1$ be the integer of Hypothesis 2.4 for $\gamma/2$, and ASSUME: an algorithm $\mathcal{A}$ solves the problem in the statement in time $O(N^{2-\gamma})$ on instances of size N of the stated form. PROVE: $\mathsf{OV}_{n,m}$ with $m = \lceil c \log n \rceil$ can be solved in time $O(n^{2-\gamma/2})$, which contradicts the choice of c.

⟨1⟩2. LET: (A, B) be an instance of $\mathsf{OV}_{n,m}$. If $0^m \in A$, the answer is YES. Otherwise every $a \in A$ has $\langle a, \mathbf{1}\rangle \geq 1$, where $\mathbf{1}$ is the all-ones vector.

PROOF: The zero vector is orthogonal to every vector of B.

⟨1⟩3. LET: $N = 2n$, and LET: n_2 be such that $r(N') \leq N'$ and $w(N') \leq \frac{1}{2}\log N'$ for all $N' \geq n_2$; inputs with $2n < n_2$ are solved by brute force in time $O(1)$, so assume $N \geq n_2$. LET: $r = r(N)$, $w = w(N)$, $K_0 = \lceil m/w \rceil$, and partition $[m]$ into consecutive blocks $B_1, \ldots, B_{K_0}$ of size at most w. For $a, b' \in \{0,1\}^m$ LET: $F(a, b')$ be the number of blocks t with $a_k b'_k = 1$ for some $k \in B_t$. Then $F(a, b') = \langle \phi(a), \psi(b')\rangle$, where $\phi(a)$ and $\psi(b')$ are indexed by the pairs (t, S) with $\emptyset \neq S \subseteq B_t$ and

$$\phi(a)_{t,S} = (-1)^{|S|+1} \prod_{k \in S} a_k, \qquad \psi(b')_{t,S} = \prod_{k \in S} b'_k.$$

They have $d_0 = \sum_t (2^{|B_t|} - 1) \leq K_0(2^w - 1)$ coordinates, with entries in $\{-1, 0, 1\}$ and $\{0, 1\}$.
PROOF: For each block, $1 - \prod_{k \in B_t}(1 - a_k b'_k)$ is 1 if a and b' share a one in B_t and 0 otherwise, and expanding the product gives $\sum_{\emptyset \neq S \subseteq B_t} (-1)^{|S|+1} \prod_{k \in S} a_k b'_k$. Sum over t.

⟨1⟩4. LET: $z = 1 + 1/r$, $\lambda = \ln z$, $\varepsilon_0 = r^{K_0}/(32n(r+1)^{2K_0})$, and LET: λ' be a rational with $\lambda - \varepsilon_0/(12K_0) \leq \lambda' \leq \lambda$. For a sublist B' of B LET: the instance $I(B')$ have the queries $q_a = \lambda' \phi(a)$ for $a \in A$ and n zero queries; for each $b' \in B$ a key equal to $\psi(b')$ if $b' \in B'$ and to $\psi(\mathbf{1})$ otherwise, with value 0; and n keys equal to 0 with value 1. Then $\lambda' > 0$.

PROOF: λ' exists with $O(\log(K_0/\varepsilon_0)) = O(K_0 \log(r+1) + \log n) = O(\log^2 n)$ bits, because $K_0 \leq m$ and $r \leq N$; this is $O(\log(N/\varepsilon))$, as the model allows. It is positive because $\lambda = \ln(1 + 1/r) \geq 1/(r+1)$, while $\varepsilon_0/(12K_0) \leq 1/(384n(r+1))$ since $r^{K_0} \leq (r+1)^{K_0}$. $I(B')$ has N queries, N keys, dimension d_0, $d_v = 1$ and values in $\{0, 1\}$. Since $w \leq \frac{1}{2}\log N$, $2^w \leq \sqrt{2n}$ and $d_0 \leq m\sqrt{2n}$, so $I(B')$ is computed in time $O(n^{3/2}\mathrm{poly}(\log n))$.

⟨1⟩5. $I(B')$ is an instance of the problem in the statement for size N.

PROOF: $K_0 \leq \lceil \lceil c \log N \rceil / w \rceil = K$, so $d_0 \leq K(2^w - 1)$ and padding with zero coordinates reaches dimension d. A query q_a has logits $\lambda' F(a, \cdot) \in [0, \lambda' K_0] \subseteq [0, W]$ for the keys $\psi(\cdot)$ and 0 for the zero keys; zero queries have all logits 0. The precision is handled in ⟨1⟩7.

⟨1⟩6. LET: o_a be the output of q_a when λ' is replaced by λ, LET: $S_a = \sum_{b' \in B} z^{F(a, \tilde{b}')}$, where $\tilde{b}'$ is b' if $b' \in B'$ and $\mathbf{1}$ otherwise, and LET: $N_a = |\{b' \in B' : \langle a, b'\rangle = 0\}|$. Then $o_a = n/(n + S_a)$, the number $X_a = r^{K_0} S_a$ is an integer, and $X_a \equiv r^{K_0} N_a \pmod{r+1}$.

⟨2⟩1. $o_a = n/(n + S_a)$.

PROOF: The n zero keys have logit 0 and value 1; the other keys have value 0 and weight $e^{\lambda F(a, \tilde{b}')} = z^{F(a, \tilde{b}')}$.

⟨2⟩2. For $0 \leq s \leq K_0$, $r^{K_0} z^s = (r+1)^s r^{K_0 - s}$ is an integer, divisible by $r+1$ if $s \geq 1$, and equal to r^{K_0} if $s = 0$.

PROOF: $z^s = (r+1)^s / r^s$, and F takes values in $\{0, \ldots, K_0\}$.

⟨2⟩3. $F(a, \tilde{b}') = 0$ holds exactly for the N_a kept vectors $b' \in B'$ with $\langle a, b'\rangle = 0$.

PROOF: $F(a, b') = 0$ if and only if a and b' share no one, that is, $\langle a, b'\rangle = 0$. For a dropped b', $F(a, \mathbf{1}) \geq 1$ by ⟨1⟩2.

⟨2⟩4. Q.E.D.

PROOF: Sum ⟨2⟩2 over the keys and use ⟨2⟩3.

⟨1⟩7. ASSUME: $\|\tilde{o} - o'\|_\infty \leq \varepsilon$ for the true outputs o' of $I(B')$, with ε as in the statement at size N, and LET: $\hat{o}_a$ be the multiple of $\varepsilon_0/8$ nearest to $\min\{1, \max\{\varepsilon_0, \tilde{o}_a\}\}$. Then $X_a = \lfloor r^{K_0} n(1/\hat{o}_a - 1) + 1/2 \rfloor$ for every $a \in A$.

⟨2⟩1. $\varepsilon \le \varepsilon_0/2$.

PROOF: $\varepsilon = 1/(64N(r+1)^{2K})$ and $\varepsilon_0/2 \ge 1/(64n(r+1)^{2K_0})$, since $r^{K_0} \ge 1$, $n \le N$ and $K_0 \le K$.

⟨2⟩2. $|o'_a - o_a| \le \varepsilon_0/4$.

PROOF: The logits of q_a change by at most $\eta = K_0(\lambda - \lambda') \le \varepsilon_0/12$, and Lemma 2.3 gives $e^{2\eta} - 1 \le 3\eta$ because $\eta \le 1/4$.

⟨2⟩3. LET: $E = z^{K_0}$. Then $o_a \ge 1/(2E) \ge 2\varepsilon_0$ and $|\hat{o}_a - o_a| \le \varepsilon_0$.

PROOF: $S_a \le nE$ gives $o_a \ge 1/(1+E) \ge 1/(2E)$, and $1/(2E) = r^{K_0}/(2(r+1)^{K_0}) \ge 2\varepsilon_0$. By ⟨2⟩1 and ⟨2⟩2, $|\tilde{o}_a - o_a| \le 3\varepsilon_0/4$. Since $o_a \in [\varepsilon_0, 1]$, replacing $\tilde{o}_a$ by $\min\{1, \max\{\varepsilon_0, \tilde{o}_a\}\}$ does not increase its distance to o_a, and rounding to a multiple of $\varepsilon_0/8$ adds at most $\varepsilon_0/16$.

⟨2⟩4. $|n(1/\hat{o}_a - 1) - S_a| \le 1/(4r^{K_0})$.

PROOF: By ⟨2⟩3, $\hat{o}_a \ge o_a - \varepsilon_0 \ge o_a/2$, and $S_a = n/o_a - n$ by ⟨1⟩6. So $|n/\hat{o}_a - n/o_a| = n|o_a - \hat{o}_a|/(o_a\hat{o}_a) \le 2n\varepsilon_0/o_a^2 \le 8n\varepsilon_0 E^2 = 8n\varepsilon_0(r+1)^{2K_0}/r^{2K_0} = 1/(4r^{K_0})$.

⟨2⟩5. Q.E.D.

PROOF: Multiply ⟨2⟩4 by r^{K_0}: the real number $r^{K_0}n(1/\hat{o}_a - 1)$ is within $1/4$ of the integer X_a.

⟨1⟩8. LET: the algorithm run $L = 4$ rounds. In each round it draws B' by keeping every $b' \in B$ independently with probability $1/2$, runs $\mathcal{A}$ on $I(B')$ independently $R = \lceil 18 \ln 40 \rceil$ times, takes the coordinatewise median $\tilde{o}$, computes $\hat{o}_a$ and X_a as in ⟨1⟩7 and $\rho_a = X_a \cdot (r^{K_0})^{-1} \bmod (r+1)$ for every a. It answers YES if some $\rho_a \ne 0$ in some round, and NO otherwise. It errs with probability at most $1/3$.

⟨2⟩1. In each round, with probability at least $39/40$, the median satisfies the hypothesis of ⟨1⟩7, and then $\rho_a = N_a \bmod (r+1)$ for all a.

PROOF: Each run satisfies $\|\cdot - o'\|_\infty \le \varepsilon$ with probability at least $2/3$. By Hoeffding's inequality, at least half of the runs fail with probability at most $e^{-R/18} \le 1/40$. Otherwise, in every coordinate more than half of the values lie within ε of o', and so does their median. Then ⟨1⟩7 and ⟨1⟩6 give $X_a \equiv r^{K_0}N_a$, and r^{K_0} is invertible modulo $r+1$ since $\gcd(r, r+1) = 1$.

⟨2⟩2. If (A, B) has no orthogonal pair, the algorithm answers NO with probability at least $9/10$.

PROOF: All $N_a = 0$; by ⟨2⟩1, with probability at least $1 - 4/40$ every ρ_a in every round is 0.

⟨2⟩3. If $\langle a^*, b^* \rangle = 0$ for some $a^* \in A$ and $b^* \in B$, then in each round $N_{a^*} \not\equiv 0 \pmod{r+1}$ with probability at least $1/2$, independently of the other rounds.

PROOF: Condition on the choices for all $b' \ne b^*$. The two possible values of N_{a^*} differ by 1, and at most one of them is divisible by $r+1 \ge 2$.

⟨2⟩4. Q.E.D.

PROOF: By ⟨2⟩2, the error on NO instances is at most $1/10$. On YES instances, by ⟨2⟩3 no round has $N_{a^*} \not\equiv 0$ with probability at most $1/16$, and by ⟨2⟩1 some median is inaccurate with probability at most $1/10$; otherwise $\rho_{a^*} \ne 0$ in some round. The error is at most $1/16 + 1/10 < 1/3$.

⟨1⟩9. The algorithm runs in time $O(n^{2-\gamma/2})$.

PROOF: Computing $r(N)$ and $w(N)$ takes time $O(n)$. Building the L instances takes time $O(n^{3/2}\mathrm{poly}(\log n))$ by ⟨1⟩4. The algorithm makes $LR = O(1)$ calls of $\mathcal{A}$ on instances of size $N = 2n$, each in time $O(n^{2-\gamma})$. Reading their outputs, taking the medians by comparisons, and truncating each median to $O(\log(1/\varepsilon_0))$ bits before rounding cost $O(R)$ times the length

of the outputs, which is at most the running time of the calls. Computing λ' to $O(\log^2 n)$ bits takes $\mathrm{poly}(\log n)$ time. Each $\hat{o}_a$ is a multiple of $\varepsilon_0/8$ in $[\varepsilon_0, 1+\varepsilon_0]$, so the X_a and ρ_a take $n\mathrm{poly}(\log n)$ operations on numbers of $O(K_0 \log(r+1) + \log n) = O(\log^2 n)$ bits. The total is $O(n^{2-\gamma} + n^{3/2}\mathrm{poly}(\log n)) = O(n^{2-\gamma/2})$, because $\gamma < 1$.

⟨1⟩10. Q.E.D.

PROOF: ⟨1⟩8 and ⟨1⟩9 contradict the choice of c in ⟨1⟩1.

□

C.2 The polynomial method

Proposition 4.4 (Polynomial method)**.** *For all $C \geq 1$, $\beta \geq 1$ and $\eta \in (0,1)$ there is $\kappa = \kappa(C, \beta, \eta) > 0$ such that the following holds for all sufficiently large n. Every attention instance with n queries, n keys, dimension $d \leq C \log n$, $d_v \leq d$ value coordinates and spread at most $\kappa \ln n$ can be solved to precision $n^{-\beta}$ deterministically with $O(n^{1+\eta})$ arithmetic operations on rationals of $O(\log^2 n)$ bits.*

Proof. ⟨1⟩1. LET: $A_0 = C/\ln 2 + 3$, choose $\gamma_0 \in (0,1]$ with $\gamma_0 \ln(eA_0/\gamma_0) \leq \eta/2$, and LET: $\kappa = \min\{1/4, (\gamma_0/(2e))\exp(-(\beta+4)/\gamma_0)\}$, $R = 2\kappa \ln n$ and $g = \lceil \gamma_0 \ln n \rceil$.

PROOF: $\gamma \ln(eA_0/\gamma) \to 0$ as $\gamma \to 0$, so γ_0 exists.

⟨1⟩2. Appending the coordinate $-\langle q_i, k_1 \rangle$ to each q_i and the coordinate 1 to each key gives an instance with the same outputs, dimension at most $C \log n + 1$, and all logits in $[-\kappa \ln n, \kappa \ln n]$.

PROOF: The logits of row i become $\ell_{ij} - \ell_{i1}$, a shift that does not change π_{ij}, and they lie in $[-\kappa \ln n, \kappa \ln n]$ because the spread is at most $\kappa \ln n$.

⟨1⟩3. LET: every entry of the shifted queries, of the keys and of the values be rounded to the nearest multiple of 2^{-L}, with $L = \lceil K_1 \log n \rceil$ for a constant K_1 depending on β and on the bit length of the input. For large n this changes every logit by at most $n^{-\beta-3}$, moves every output by at most $4n^{-\beta-3}$, and afterwards every entry is an integer multiple of 2^{-L} of absolute value at most n^{K_1}.

PROOF: The input entries are rationals of $O(\log n)$ bits, hence at most n^{K_2} in absolute value for a constant K_2, and so are the shifted entries. A logit is a sum of at most $C \log n + 1$ products, each of which changes by at most $2n^{K_2}2^{-L} + 2^{-2L}$; with $K_1 \geq K_2 + \beta + 5$ the total is at most $n^{-\beta-3}$. Lemma 2.3 bounds the effect of the logits by $e^{2n^{-\beta-3}} - 1 \leq 3n^{-\beta-3}$, and rounding the values moves each output by at most $2^{-L} \leq n^{-\beta-3}$.

⟨1⟩4. LET: $P_g(x) = \sum_{t<g} x^t/t!$. For $|x| \leq R$, $|P_g(x)/e^x - 1| \leq e^{2R}(eR/g)^g \leq n^{-\beta-2}$ for large n.

PROOF: By Taylor's theorem with Lagrange remainder, $|e^x - P_g(x)| \leq e^R R^g/g!$; divide by $e^x \geq e^{-R}$ and use $g! \geq (g/e)^g$. Then $eR/g \leq 2e\kappa/\gamma_0 \leq \exp(-(\beta+4)/\gamma_0)$, so $(eR/g)^g \leq n^{-\beta-4}$, and $e^{2R} = n^{4\kappa} \leq n$. All rounded logits lie in $[-R, R]$, because $\kappa \ln n + n^{-\beta-3} \leq R$.

⟨1⟩5. LET: $\Phi_q(x) = (x^\alpha/\alpha!)_{|\alpha|<g}$ and $\Phi_k(x) = (x^\alpha)_{|\alpha|<g}$, indexed by multi-indices α with $|\alpha| < g$. Then $P_g(\langle q, k \rangle) = \langle \Phi_q(q), \Phi_k(k) \rangle$, and both vectors have $r_g = \binom{d'+g-1}{d'} \leq \binom{d'+g}{g}$ coordinates, where $d' \leq C \log n + 1$ is the dimension after ⟨1⟩2.

PROOF: The multinomial theorem gives $\langle q, k \rangle^t/t! = \sum_{|\alpha|=t} \prod_u (q_u k_u)^{\alpha_u}/\alpha_u!$.

⟨1⟩6. $r_g \leq (eA_0/\gamma_0)\, n^{\eta/2}$ for all large n.

PROOF: $\binom{d'+g}{g} \leq (e(d'+g)/g)^g$, and $(d'+g)/g \leq ((C/\ln 2)\ln n + 1 + \gamma_0 \ln n + 1)/(\gamma_0 \ln n) \leq A_0/\gamma_0$ for large n. So $r_g \leq (eA_0/\gamma_0)^{\gamma_0 \ln n + 1} = (eA_0/\gamma_0)\, n^{\gamma_0 \ln(eA_0/\gamma_0)}$.

⟨1⟩7. LET: the algorithm compute $\Phi_q(q_i)$ and $\Phi_k(k_j)$ for the rounded vectors, then $U = \sum_j \Phi_k(k_j) v_j^\top$ and $u = \sum_j \Phi_k(k_j)$, and output $\tilde{o}_i = U^\top \Phi_q(q_i)/\langle u, \Phi_q(q_i) \rangle$. It uses $O(n r_g(d' + d_v)) = O(n^{1+\eta})$ arithmetic operations on rationals of $O(\log^2 n)$ bits.

PROOF: Each monomial of degree $t \geq 1$ is a monomial of degree $t-1$ times one coordinate, so each vector takes $O(r_g)$ operations after an $O(r_g d')$ enumeration of the multi-indices; the products and sums take $O(nr_g d_v)$ operations, and by $\langle 1\rangle 6$ the total is $O(n^{1+\eta})$. By $\langle 1\rangle 3$, a monomial of degree below g is an integer of $O(g \log n)$ bits divided by $2^{L(g-1)}$; with the common denominator $2^{L(g-1)}(g-1)!$ for Φ_q, and 2^L for the values, every number in the computation has a numerator of $O(g \log n + \log(nr_g)) = O(\log^2 n)$ bits.

$\langle 1\rangle 8$. Q.E.D.

PROOF: By $\langle 1\rangle 5$, $\langle u, \Phi_q(q_i)\rangle = \sum_j P_g(\ell'_{ij})$ and $U^\top \Phi_q(q_i) = \sum_j P_g(\ell'_{ij}) v'_j$ for the rounded logits ℓ' and values v', so $\tilde{o}_i$ is the average of the v'_j with weights $P_g(\ell'_{ij})$. By $\langle 1\rangle 4$ these weights are $e^{\ell'_{ij}}$ times factors in $[1-n^{-\beta-2}, 1+n^{-\beta-2}]$, which moves the average by at most $3n^{-\beta-2}$, as in Lemma 2.3. With $\langle 1\rangle 3$, the total error is at most $3n^{-\beta-2} + 4n^{-\beta-3} \leq n^{-\beta}$.

□

C.3 Dimension matters at high precision

Proposition 4.6 (Polynomial method in higher dimension)**.** *Let $C > 1$, $\beta \geq 1$ and $\eta \in (0,1)$ with $4(\beta+5)(C-1) < \eta$, and let $a \in (4(\beta+5)(C-1)/\eta, 1]$. For all sufficiently large n, every attention instance with n queries, n keys, dimension $d \leq (\log n)^C$, $d_v \leq d$ value coordinates and spread at most $(\log n)^{1-a}$ can be solved to precision $n^{-\beta}$ deterministically with $O(n^{1+\eta})$ arithmetic operations on rationals of $O(\log^2 n)$ bits.*

Proof. $\langle 1\rangle 1$. LET: $\gamma_1 = 2(\beta+5)/a$, $R = (\log n)^{1-a} + 1$ and $g = \lceil \gamma_1 \ln n / \ln\ln n \rceil$, and assume n is large.

$\langle 1\rangle 2$. Steps $\langle 1\rangle 2$, $\langle 1\rangle 3$, $\langle 1\rangle 5$, $\langle 1\rangle 7$ and $\langle 1\rangle 8$ of the proof of Proposition 4.4 hold with $d' \leq (\log n)^C + 1$, with R and g as in $\langle 1\rangle 1$, with the spread bound $(\log n)^{1-a}$ in place of $\kappa \ln n$, and with steps $\langle 1\rangle 2$ and $\langle 1\rangle 3$ there carried out in the opposite order, provided that $|P_g(x)/e^x - 1| \leq n^{-\beta-2}$ for $|x| \leq R$ and $r_g \leq n^{\eta/2}$.

PROOF: First every entry of the queries, keys and values is rounded to the nearest multiple of 2^{-L}, as in step $\langle 1\rangle 3$ there; the input entries are rationals of $O(\log n)$ bits, hence at most n^{K_2} in absolute value. A logit is a sum of at most $(\log n)^C$ products, and $(\log n)^C(2n^{K_2-K_1} + n^{-2K_1}) \leq n^{-\beta-3}$ for large n when $K_1 \geq K_2 + \beta + 5$; so every logit moves by at most $n^{-\beta-3}$ and every output by at most $4n^{-\beta-3}$. Then the coordinate $-\langle \tilde{q}_i, \tilde{k}_1\rangle$ of the rounded vectors is appended to each query and 1 to each key, as in step $\langle 1\rangle 2$ there. This coordinate is a multiple of 2^{-2L} of absolute value $n^{O(1)}$, the shift preserves the outputs of the rounded instance, and every shifted logit lies within $(\log n)^{1-a} + 2n^{-\beta-3} \leq R$ of 0. Step $\langle 1\rangle 5$ holds in every dimension. Step $\langle 1\rangle 7$ uses $O(nr_g(d'+d_v)) = O(n^{1+\eta/2}(\log n)^C) = O(n^{1+\eta})$ operations; with the grid 2^{-2L} in place of 2^{-L}, a monomial of degree below g is an integer of $O(g \log n)$ bits divided by $2^{2L(g-1)}$, so every number has $O(g \log n + \log(nr_g)) = O(\log^2 n)$ bits. Step $\langle 1\rangle 8$ uses the Taylor bound of the proviso, the containment of the shifted rounded logits in $[-R, R]$, and the rounding bound.

$\langle 1\rangle 3$. For $|x| \leq R$, $|P_g(x)/e^x - 1| \leq e^{2R}(eR/g)^g \leq n^{-\beta-2}$.

PROOF: The first inequality is shown in step $\langle 1\rangle 4$ of the proof of Proposition 4.4. Since $a > 0$, $e^{2R} = e^{2(\log n)^{1-a}+2} \leq n$ for large n. Since $a \leq 1$, $R \leq 2(\log n)^{1-a}$, and with $\ln n = \ln 2 \cdot \log n$,

$$\frac{g}{eR} \geq \frac{\gamma_1 \ln 2}{2e} \cdot \frac{(\log n)^a}{\ln\ln n}.$$

As $\ln\log n \geq \ln\ln n$, this gives $\ln(g/(eR)) \geq a\ln\ln n - \ln\ln\ln n + \ln(\gamma_1 \ln 2/(2e)) \geq (a/2)\ln\ln n$ for large n, and so $g\ln(g/(eR)) \geq (\gamma_1 \ln n/\ln\ln n)(a/2)\ln\ln n = (\beta+5)\ln n$. Hence $(eR/g)^g \leq n^{-\beta-5}$ and $e^{2R}(eR/g)^g \leq n^{-\beta-4}$.

⟨1⟩4. $r_g \le n^{\eta/2}$.

PROOF: $r_g = \binom{d'+g-1}{g-1} \le (e(d'+g-1)/(g-1))^{g-1}$. For large n, $g-1 \ge \gamma_1 \ln n/(2\ln\ln n)$ and $d' \le 2(\log n)^C$, so

$$\frac{d'+g-1}{g-1} = 1 + \frac{d'}{g-1} \le 1 + \frac{4(\log n)^C \ln\ln n}{\gamma_1 \ln n} \le \Big(1+\frac{6}{\gamma_1}\Big)(\log n)^{C-1}\ln\ln n.$$

With $g-1 \le \gamma_1 \ln n/\ln\ln n$ and $\ln\log n \le \ln\ln n + 1$,

$$\ln r_g \le \frac{\gamma_1 \ln n}{\ln\ln n}\Big((C-1)(\ln\ln n+1) + \ln\ln\ln n + 1 + \ln(1+6/\gamma_1)\Big) = \gamma_1(C-1)\ln n + o(\ln n).$$

Since $\gamma_1(C-1) = 2(\beta+5)(C-1)/a < \eta/2$ by the choice of a, $r_g \le n^{\eta/2}$ for large n.

⟨1⟩5. Q.E.D.

PROOF: By ⟨1⟩2, ⟨1⟩3 and ⟨1⟩4.

□

D Proofs for Section 5

Imported Theorem 2.7 (Williams [Wil05])**.** *A CNF formula with N variables, N even, and M clauses can be transformed in time $2^{N/2}\text{poly}(N,M)$ into an instance of $\mathsf{OV}_{n,M}$ with $n = 2^{N/2}$ that has an orthogonal pair if and only if the formula is satisfiable.*

Proof. ⟨1⟩1. LET: the variables be split into halves X and Y. For $x \in \{0,1\}^X$ LET: $a_x \in \{0,1\}^M$ have $a_x[C] = 1$ if and only if x satisfies no literal of clause C over X; define b_y for $y \in \{0,1\}^Y$ in the same way.

⟨1⟩2. (x,y) satisfies the formula if and only if $\langle a_x, b_y\rangle = 0$.

PROOF: (x,y) falsifies C if and only if x satisfies no literal of C over X and y none over Y, that is, $a_x[C] = b_y[C] = 1$.

⟨1⟩3. Q.E.D.

PROOF: The lists $(a_x)_x$ and $(b_y)_y$ have $2^{N/2}$ entries each, repetitions included, and are computed in time $2^{N/2}\text{poly}(N,M)$.

□

D.1 CNF-SETH in polylogarithmic dimension

Proposition 5.1 (CNF-SETH, exact embedding)**.** *Assume Hypothesis 2.5. For every $\gamma \in (0,1)$ there is $C \ge 1$ such that, with $d = \lceil(\log n)^C\rceil$, the problem $\mathsf{ATT}(n,d,d\ln(16n),1/6)$ cannot be solved in time $O(n^{2-\gamma})$. This holds already for $d_v = 1$ and values in $\{0,1\}$.*

Proof. ⟨1⟩1. LET: c_0 be the constant of Hypothesis 2.5 for $\gamma/2$, and $C = c_0 + 1$. ASSUME: an algorithm $\mathcal{A}$ solves $\mathsf{ATT}(n, \lceil(\log n)^C\rceil, \lceil(\log n)^C\rceil\ln(16n), 1/6)$ in time $O(n^{2-\gamma})$. PROVE: CNF formulas with N variables and at most N^{c_0} clauses can be decided in time $O(2^{(1-\gamma/2)N})$.

⟨1⟩2. LET: F have N variables, N even after adding a dummy variable, and $M \le N^{c_0}$ clauses, and LET: (A,B) be the instance of $\mathsf{OV}_{n,M}$ of Imported Theorem 2.7, $n = 2^{N/2}$. Then $M \le (2\log n)^{c_0}$.

⟨1⟩3. LET: $\lambda = \lceil 10\ln(14n)\rceil/10$. LET: the instance I have the queries $q_a = (-\lambda a, -11/10)$ for $a \in A$ and one zero query, the keys $k_b = (b,0)$ with value 1 for $b \in B$, and the key $k^* = (0,1)$ with

value 0. Its size is $n+1$, its dimension $M+1 \le \lceil (\log(n+1))^C \rceil$ for large n, and its spread at most $(M+1)\ln(16n)$.

PROOF: The logits of q_a are $-\lambda\langle a,b\rangle \in [-\lambda M, 0]$ and $-11/10$, and $\lambda M \le M(\ln(14n)+1/10) \le M\ln(16n)$, as $\ln(16/14) > 1/10$. The zero query has all logits 0. Finally $(2\log n)^{c_0}+1 \le (\log n)^{c_0+1}$ for large n.

⟨1⟩4. If a has an orthogonal partner in B, its output is at least $3/4$; otherwise at most $1/4$.

PROOF: The output is $S_a/(S_a+e^{-11/10})$ with $S_a = \sum_b e^{-\lambda\langle a,b\rangle}$. An orthogonal partner gives $S_a \ge 1$ and output at least $1/(1+e^{-11/10}) > 3/4$. Otherwise $S_a \le ne^{-\lambda} \le ne^{-\ln(14n)}$, and the output is at most $S_a e^{11/10} \le e^{11/10}/14 < 1/4$.

⟨1⟩5. Q.E.D.

PROOF: Run $\mathcal{A}$ once on I and answer YES if and only if some q_a receives an output above $1/2$. By ⟨1⟩4 and Imported Theorem 2.7 the answer is correct with probability at least $2/3$. The time is $2^{N/2}\text{poly}(N) + O((n+1)^{2-\gamma}) = O(2^{(1-\gamma/2)N})$, contradicting the choice of c_0.

□

D.2 Clause exponent below two

Lemma 5.2 (Gap instances from MA protocols)**.** *Let $m \ge s \ge 2$ and $k \ge 1$ be integers, and let p, ρ and ℓ be the lengths of Merlin's proof, of the coins and of Bob's message in the protocol of Imported Theorem 2.9 for (m,s). There is a map from instances (A,B) of $\mathsf{OV}_{n,m}$ to 2^p instances of Boolean Max-IP, each with n vectors per side in dimension $2^{k(\rho+\ell)}$, such that with $T_k = 2^{k\rho}$:*

(i) every inner product lies in $[0,T_k]$;

(ii) if (A,B) has an orthogonal pair, some instance has a pair with inner product T_k;

(iii) otherwise every inner product is at most $2^{-k}T_k$.

The map takes time $2^{p+k(\rho+\ell)}\,nk\text{poly}(m)$.

Proof. ⟨1⟩1. Pad Merlin's proof, the coins and Bob's message of the protocol of Imported Theorem 2.9 to exactly p, ρ and ℓ bits, and LET: $\nu(b,w)$ be Bob's message on input b and coins w. LET: the k-fold protocol use coins $w = (w_1,\ldots,w_k) \in \{0,1\}^{k\rho}$, LET: Bob send $\nu_k(b,w) = (\nu(b,w_1),\ldots,\nu(b,w_k))$, and LET: Alice accept on proof μ if and only if she accepts every pair $(w_i,\nu(b,w_i))$ in the original protocol. Then disjoint inputs have a proof that is accepted for all coins, and for intersecting inputs every proof is accepted with probability at most 2^{-k}.

PROOF: Unused coins do not change acceptance probabilities. For a fixed proof, the k checks use independent coins, so the acceptance probabilities multiply.

⟨1⟩2. For $\mu \in \{0,1\}^p$ and $a \in A$ LET: $x_{\mu,a} \in \{0,1\}^{\{0,1\}^{k\rho}\times\{0,1\}^{k\ell}}$ have $x_{\mu,a}[w,\zeta] = 1$ if and only if Alice accepts on input a, proof μ, coins w and message ζ. For $b \in B$ LET: $y_b[w,\zeta] = 1$ if and only if $\zeta = \nu_k(b,w)$. Instance μ is $((x_{\mu,a})_{a\in A},(y_b)_{b\in B})$.

⟨1⟩3. $\langle x_{\mu,a}, y_b\rangle$ equals the number of coin strings w on which Alice accepts a, b and μ.

PROOF: For each w, the vector y_b has exactly one 1 in block w, at $\zeta = \nu_k(b,w)$.

⟨1⟩4. Q.E.D.

PROOF: By ⟨1⟩3, every inner product is an integer in $[0,2^{k\rho}]$, which is (i). By ⟨1⟩1, an orthogonal pair (a,b) has a proof μ with inner product $2^{k\rho}$, which is (ii), and if there is no orthogonal pair every inner product is at most $2^{-k}2^{k\rho}$, which is (iii). Each of the $2^{k(\rho+\ell)}$ coordinates of each of the $(2^p+1)n$ vectors needs k evaluations of the protocol's predicates, each computable in time $\text{poly}(m)$.

□

Theorem 5.3 (Clause exponent below two)**.** *Assume* CNF-SETH$_{c_0}$ *for some* $c_0 < 2$*. For all* $\gamma \in (0,1)$*,* $\theta \in (0,1]$ *and* $\varepsilon \in (0,1/2)$ *there is* $K \geq 1$ *such that* $\mathsf{ATT}(n, d(n), (1+\theta)\ln n, \varepsilon)$ *with* $d(n) = \lceil 2^{K(\log n)^{c_0-1}(\log\log n)^2} \rceil = n^{o(1)}$ *cannot be solved in time* $O(n^{2-\gamma})$*. This holds already for instances with* $d_v = 1$*, values in* $\{0,1\}$*, and all logits in* $[0,(1+\theta)\ln n]$*.*

Proof. ⟨1⟩1. Let: $D = \lceil (2\log n)^{c_0} \rceil$. Then $\mathsf{OV}_{n,D}$ cannot be solved in time $O(n^{2-\gamma/2})$.
Proof: An algorithm in time $O(n^{2-\gamma/2})$ would decide CNF formulas with N variables and at most N^{c_0} clauses in time $O(2^{(1-\gamma/4)N})$ by Imported Theorem 2.7, contradicting CNF-SETH$_{c_0}$ for $\gamma/4$.

⟨1⟩2. Let: $k = \lceil \log(4/\theta) \rceil$ and $s = \lceil 8a_1 D \log D/(\gamma \log n) \rceil$. For large n, $2 \leq s \leq D$, Merlin's proof has $p \leq (\gamma/8)\log n$ bits, and $k(\rho+\ell) \leq K_0(\log n)^{c_0-1}(\log\log n)^2$ for a constant K_0 depending on γ, θ and c_0.
Proof: $D \geq 2\log n$, so $D\log D/\log n \to \infty$ and $\log D = O(\log\log n)$; hence $2 \leq s \leq D$ for large n. By Imported Theorem 2.9, $p \leq a_1(D/s)\log s \leq (\gamma \log n/(8\log D))\log s \leq (\gamma/8)\log n$. Also $\rho \leq a_2 \log D$ and $\ell \leq a_3 s \log s = O(D\log^2 D/(\gamma\log n)) = O((\log n)^{c_0-1}(\log\log n)^2/\gamma)$.

⟨1⟩3. Let: $K = K_0 + 1$. For each of the 2^p instances of Lemma 5.2 with these parameters, build the attention instance of step ⟨1⟩3 in the proof of Theorem 3.3, with T_k in place of T and 2^k in place of τ. It has $n+1$ queries and keys, dimension $2^{k(\rho+\ell)} + 1 \leq d(n+1)$, all logits in $[0, \lambda T_k]$ with $\lambda T_k \leq (1+\theta)\ln n$ for large n, and a query q_x has output at least $1-\eta$ if x has a partner at inner product T_k and at most η if all its inner products are at most $2^{-k}T_k$, where $\eta = (1-2\varepsilon)/4$.
Proof: Steps ⟨1⟩4 and ⟨1⟩5 of that proof use only that all inner products lie in $[0,T]$, that an instance has at most n real keys, and that $1/\tau \leq \theta/4$; these hold with T_k and $\tau = 2^k \geq 4/\theta$ by Lemma 5.2(i).

⟨1⟩4. Assume: an algorithm $\mathcal{A}$ solves $\mathsf{ATT}(n, d(n), (1+\theta)\ln n, \varepsilon)$ in time $O(n^{2-\gamma})$. Then $\mathsf{OV}_{n,D}$ can be solved in time $O(n^{2-\gamma/2})$.
Proof: As in step ⟨1⟩6 of the proof of Theorem 3.3, with $N_\psi = 2^p \leq n^{\gamma/8}$ instances and $R = \lceil 18\ln(3\cdot 2^p) \rceil$ runs each; that step uses the two implications of its step ⟨1⟩5, which hold here by ⟨1⟩3, and items (ii) and (iii) of Imported Theorem 2.8, which hold here by Lemma 5.2(ii) and (iii). The calls take $2^p R \cdot O((n+1)^{2-\gamma}) = O(n^{2-7\gamma/8}\log n)$, and Lemma 5.2 takes $2^{p+k(\rho+\ell)} nk\,\mathrm{poly}(D) = n^{1+\gamma/8+o(1)}$, since $c_0 < 2$ makes $k(\rho+\ell) = o(\log n)$.

⟨1⟩5. Q.E.D.
Proof: ⟨1⟩4 contradicts ⟨1⟩1. Finally $d(n) = n^{o(1)}$ because $(\log n)^{c_0-1}(\log\log n)^2 = o(\log n)$ for $c_0 < 2$.

□

D.3 A barrier for the bounded-entry threshold

The barrier rests on the following tradeoff for Merlin–Arthur protocols, which is Klauck's argument [Kla03, Theorem 1, part 1] with the proof length and the communication kept apart. In a *public-coin MA protocol* for a function f, Merlin sends a proof $z \in \{0,1\}^{c_M}$ to both players, who then run a public-coin randomized protocol whose transcript has at most $c_P \geq 1$ bits and determines the output. The protocol computes f with error $1/4$ if every input with $f = 1$ has a proof accepted with probability at least $3/4$, and every input with $f = 0$ is accepted with probability at most $1/4$ under every proof. The one-message protocols of Section 2.3 are of this form after Alice announces her decision.

Lemma D.1 (Klauck's tradeoff)**.** *There is an absolute constant $\alpha_0 \in (0,1]$ such that every public-coin MA protocol computing DISJ_D with error $1/4$ satisfies $(c_M+1)\,c_P \geq \alpha_0 D$.*

Proof. ⟨1⟩1. Let: $\alpha_1, \varepsilon_1, \mu$ be as in Imported Theorem 2.10, with $\alpha_1 \leq 1$ without loss of generality, Let: μ_1 and μ_0 be μ conditioned on $\mathsf{DISJ}_D^{-1}(1)$ and on $\mathsf{DISJ}_D^{-1}(0)$, and Let: $\eta = \alpha_1 2^{-c_M}/16$.

⟨1⟩2. Let: $t = \lceil 8(c_M + \log(16/\alpha_1))\rceil$. Repeating the randomized part t times with fresh coins and taking the majority gives, for every fixed proof, a protocol with transcripts of at most $k = tc_P$ bits whose acceptance probability is at least $1-\eta$ when the original one is at least $3/4$, and at most η when the original one is at most $1/4$.

Proof: By Hoeffding's inequality, the majority of t independent runs errs with probability at most $e^{-t/8} \leq \eta$, because $t/8 \geq c_M \ln 2 + \ln(16/\alpha_1)$. The t transcripts determine the majority.

⟨1⟩3. There are a proof z^* and a set $G \subseteq \mathsf{DISJ}_D^{-1}(1)$ with $\mu_1(G) \geq 2^{-c_M}$ such that the protocol of ⟨1⟩2 with proof z^* accepts every input of G with probability at least $1-\eta$.

Proof: Assign to every disjoint input a proof accepted with probability at least $3/4$. One of the at most 2^{c_M} proofs receives μ_1-mass at least 2^{-c_M}.

⟨1⟩4. There is a coin string such that the resulting deterministic protocol, with accepting set Acc, satisfies $\mu_1(G \cap \mathrm{Acc}) \geq \mu_1(G)/4$ and $\mu_0(\mathrm{Acc}) \leq 2\eta\,\mu_1(G \cap \mathrm{Acc})/\mu_1(G)$.

Proof: Over the coins, $\mathbb{E}[\mu_1(G \cap \mathrm{Acc})] \geq (1-\eta)\mu_1(G)$ by ⟨1⟩3, and $\mathbb{E}[\mu_0(\mathrm{Acc})] \leq \eta$ by ⟨1⟩2 and soundness. So $Y = \mu_1(G \cap \mathrm{Acc}) - \frac{\mu_1(G)}{2\eta}\mu_0(\mathrm{Acc})$ has $\mathbb{E}Y \geq (\frac{1}{2} - \eta)\mu_1(G) \geq \mu_1(G)/4$, and some coin string has $Y \geq \mu_1(G)/4$. Both claims follow, as $\mu_0(\mathrm{Acc}) \geq 0$.

⟨1⟩5. $k \geq \varepsilon_1 D - c_M - \log(8/\alpha_1)$.

⟨2⟩1. Acc is the disjoint union of at most 2^k rectangles.

Proof: The inputs with a given transcript form a rectangle, and the transcript determines the output.

⟨2⟩2. $\frac{1}{4}\mu_0(\mathrm{Acc}) \geq \frac{3}{4}\alpha_1\,\mu_1(G \cap \mathrm{Acc}) - 2^{k-\varepsilon_1 D}$.

Proof: Apply Imported Theorem 2.10 to each rectangle of ⟨2⟩1 and add, using $\mu(\cdot \cap \mathsf{DISJ}_D^{-1}(0)) = \frac{1}{4}\mu_0$, $\mu(\cdot \cap \mathsf{DISJ}_D^{-1}(1)) = \frac{3}{4}\mu_1$ and $\mu_1(\mathrm{Acc}) \geq \mu_1(G \cap \mathrm{Acc})$.

⟨2⟩3. Q.E.D.

Proof: By ⟨1⟩4, $\frac{1}{4}\mu_0(\mathrm{Acc}) \leq \frac{\eta}{2\mu_1(G)}\mu_1(G \cap \mathrm{Acc}) \leq \frac{\alpha_1}{32}\mu_1(G \cap \mathrm{Acc})$, since $\eta \leq \alpha_1\mu_1(G)/16$. With ⟨2⟩2, $2^{k-\varepsilon_1 D} \geq (\frac{3}{4} - \frac{1}{32})\alpha_1\mu_1(G \cap \mathrm{Acc}) \geq \frac{\alpha_1}{8}\mu_1(G) \geq \frac{\alpha_1}{8}2^{-c_M}$.

⟨1⟩6. Q.E.D.

Proof: Let: $K_1 = 8\log(16/\alpha_1) + 9$ and $K_2 = \log(8/\alpha_1)$, so that $k \leq (8c_M + K_1)c_P$ and $(8c_M + K_1)c_P \geq \varepsilon_1 D - c_M - K_2$ by ⟨1⟩5. If $c_M \geq \varepsilon_1 D/4$, then $(c_M+1)c_P \geq c_M \geq \varepsilon_1 D/4$, as $c_P \geq 1$. If $D < 4K_2/\varepsilon_1$, then $(c_M+1)c_P \geq 1 > \varepsilon_1 D/(4K_2)$. Otherwise $\varepsilon_1 D - c_M - K_2 \geq \varepsilon_1 D/2$, and $(c_M+1)c_P \geq (8c_M + K_1)c_P/K_1 \geq \varepsilon_1 D/(2K_1)$, because $K_1 \geq 8$. Take $\alpha_0 = \min\{\varepsilon_1/4, \varepsilon_1/(4K_2), \varepsilon_1/(2K_1), 1\}$.

□

Theorem 5.6 (Barrier)**.** *There is an absolute constant $\alpha_2 > 0$ such that every (D,p,d,Δ)-gap embedding with $D \geq 1$ and scale σ satisfies $\Gamma = \sigma/\Delta \geq 1/2$ and*

$$\Gamma^2 \log(2+\Gamma) \;\geq\; \alpha_2 \frac{D}{p+1}.$$

Proof. ⟨1⟩1. Let: $Q_2 = \max \|q(\mu,a)\|_2$ and $K_2 = \max \|k(\mu,b)\|_2$, so that $\sigma = Q_2 K_2$. Every logit $\langle q(\mu,a), k(\mu,b)\rangle$ lies in $[-\sigma,\sigma]$, and $\Delta \leq 2\sigma$; in particular $\sigma > 0$ and $\Gamma \geq 1/2$.

PROOF: $|\langle q,k\rangle| \le \|q\|_2\|k\|_2 \le \sigma$. With e_1 the first unit vector, (G1) for the disjoint pair $(e_1,0)$ gives μ with $\langle q(\mu,e_1),k(\mu,0)\rangle \ge h(\mu,e_1)$, and (G2) for the intersecting pair (e_1,e_1) gives $\langle q(\mu,e_1),k(\mu,e_1)\rangle \le h(\mu,e_1)-\Delta$ for the same μ. Two logits in $[-\sigma,\sigma]$ differ by at least Δ.

⟨1⟩2. LET: $u = \lceil 384\,\Gamma^2\rceil$, $t_0 = \sqrt{2\ln(32u)}$ and $\ell_0 = \lceil \log(8t_0^2\Gamma)\rceil$. LET: the MA protocol for DISJ_D on inputs a,b be: Merlin sends μ; public coins pick independent uniform $s_1,\dots,s_u \in \{-1,1\}^d$; for each i Bob sends the integer $\hat{k}_i$ nearest to $2^{\ell_0}Y_i/(t_0K_2)$, where Y_i is $\langle s_i,k(\mu,b)\rangle$ clamped to $[-t_0K_2,t_0K_2]$; Alice computes $Z = \frac{1}{u}\sum_i \langle s_i,q(\mu,a)\rangle\, t_0K_2 2^{-\ell_0}\hat{k}_i$ and sends 1 if and only if $Z \ge h(\mu,a)-\Delta/2$. It computes DISJ_D with error 1/4, with $c_M = p$ and $c_P \le 3885\,\Gamma^2\log(2+\Gamma)$.

⟨2⟩1. LET: $q = q(\mu,a)$, $k = k(\mu,b)$ and $X_i = \langle s_i,q\rangle\langle s_i,k\rangle$. The X_i are independent, have mean $\langle q,k\rangle$, and satisfy $\mathbb{E}X_i^2 \le 3\sigma^2$.

PROOF: $\mathbb{E}[s_is_i^\top]$ is the identity, which gives the mean. For uniform $s \in \{-1,1\}^d$ and $x \in \mathbb{R}^d$, $\mathbb{E}\langle s,x\rangle^4 = 3\|x\|_2^4 - 2\sum_t x_t^4 \le 3\|x\|_2^4$, so by the Cauchy–Schwarz inequality $\mathbb{E}X_i^2 \le (\mathbb{E}\langle s_i,q\rangle^4\,\mathbb{E}\langle s_i,k\rangle^4)^{1/2} \le 3\|q\|_2^2\|k\|_2^2 \le 3\sigma^2$.

⟨2⟩2. With probability at least 3/4, $|\frac{1}{u}\sum_i X_i - \langle q,k\rangle| < \Delta/4$, and $|\langle s_i,q\rangle| \le t_0Q_2$ and $|\langle s_i,k\rangle| \le t_0K_2$ for all i.

PROOF: By Chebyshev's inequality and ⟨2⟩1, the first condition fails with probability at most $3\sigma^2/(u(\Delta/4)^2) = 48\Gamma^2/u \le 1/8$. By Hoeffding's inequality, $\Pr[|\langle s,x\rangle| > t_0\|x\|_2] \le 2e^{-t_0^2/2} = 1/(16u)$ for each of the $2u$ sums, and a union bound gives 1/8.

⟨2⟩3. On the event of ⟨2⟩2, $|Z - \langle q,k\rangle| < \Delta/2$.

PROOF: No clamping occurs. So the ith term of Z differs from X_i by at most

$$|\langle s_i,q\rangle| \cdot t_0K_2\,2^{-\ell_0-1} \le t_0^2\sigma\,2^{-\ell_0-1} \le \Delta/16,$$

because $2^{\ell_0} \ge 8t_0^2\Gamma$ and $\sigma/\Gamma = \Delta$. Hence $|Z - \langle q,k\rangle| < \Delta/4 + \Delta/16$.

⟨2⟩4. Completeness and soundness hold with error 1/4.

PROOF: If $\langle a,b\rangle = 0$, (G1) gives μ with $\langle q,k\rangle \ge h(\mu,a)$, and by ⟨2⟩2 and ⟨2⟩3 Alice accepts with probability at least 3/4. If $\langle a,b\rangle \ge 1$, every μ has $\langle q,k\rangle \le h(\mu,a)-\Delta$ by (G2), and Alice accepts with probability at most 1/4. Alice knows $h(\mu,a)$, Δ, t_0, ℓ_0 and K_2.

⟨2⟩5. $c_P \le 3885\,\Gamma^2\log(2+\Gamma)$.

PROOF: Each $\hat{k}_i$ is an integer in $[-2^{\ell_0},2^{\ell_0}]$ and takes at most $\ell_0+2 \le \log(64t_0^2\Gamma)$ bits, and $c_P \le u(\ell_0+2)+1$. As $\Gamma \ge 1/2$, $u \le 388\Gamma^2$ and $t_0^2 \le 2\ln(12416\Gamma^2)$, so $64t_0^2\Gamma \le 256\Gamma\ln(112\Gamma) \le (103\Gamma)^2$, using $\ln y \le y/e$. Hence $c_P \le 776\Gamma^2\log(103\Gamma)+1 \le 777\Gamma^2\log(103\Gamma)$, as $\Gamma^2\log(103\Gamma) \ge 1$. Finally $\log(103\Gamma) \le 5\log(2+\Gamma)$, because $(2+\Gamma)^5/\Gamma$ is smallest at $\Gamma = 1/2$, where it exceeds 103.

⟨1⟩3. $c_P \ge \alpha_0 D/(p+1)$.

PROOF: By Lemma D.1 with $c_M = p$.

⟨1⟩4. Q.E.D.

PROOF: By ⟨1⟩2 and ⟨1⟩3, $3885\Gamma^2\log(2+\Gamma) \ge \alpha_0 D/(p+1)$. Take $\alpha_2 = \alpha_0/3885$.

□