Subquadratic Algorithms and Hardness for Attention with Any Temperature

Paper reading and proof discussion · 20–30 minutes

Paper authors: Shreya Gupta, Boyang Huang, Barna Saha, Yinzhan Xu, Christopher Ye

University of California, San Diego | Reader-prepared presentation

arXiv:2505.14840v1 · Read on September 28, 2026

Fixed dimension permits high-precision attention without a high-temperature restriction.

The output is small; the guarantee is strong

$$Q, K, V \in [-B, B]^{n \times d}, \qquad O = D^{-1}AV, \quad A = e^{QK^\top}, \quad D = \mathrm{diag}(A\mathbf{1}).$$

Approximate Attention Computation

Return $\widehat{O}$ with $\|\widehat{O} - O\|_\infty := \max_{i,t} |\widehat{O}_{i,t} - O_{i,t}| < \varepsilon$.

- n: context length; d: head dimension; B: entry bound.
- Explicit pairwise computation costs $O(n^2 d)$; output size is nd.
- The target is an error guarantee for *every output entry*.

Definitions 1.1–1.2, PDF p. 2. Exponentiation is entrywise.

Temperature becomes an input-magnitude parameter

$$e^{QK^\top/T} = e^{(Q/\sqrt{T})(K/\sqrt{T})^\top}, \qquad T > 0.$$

- A smaller temperature can create large, sharply separated scores.
- Seek runtime polynomial in the representation length of the magnitude: polylog B.
- With unchanged values, use $B_T = \max\{B, B/\sqrt{T}\}$.

A comparison needs a common scale

This paper uses $e^{QK^\top}$; Alman–Song uses $e^{QK^\top/d}$. Entry-bound thresholds change under this rescaling.

Introduction, PDF pp. 2–3; footnote 9, p. 17. External definition check: arXiv:2302.13214v2.

Dimension separates the main algorithmic regimes

Dimension	Main bound	Conditions
$d = O(1)$	$\widetilde{O}(n^{2-1/d}\,\mathrm{polylog}(B/\varepsilon))$	Algorithm
$d = 2^{C\log^* n}$	$n^{2-\delta}$	SETH; suitable $C(\delta)$
$d = \mathrm{poly}(n)$	$T_{\mathrm{MUL}}(n,d,n)^{1-o(1)}$	Generalized high-dim. OV hypothesis

- Small-dimension hardness: polynomial B, constant error.
- Large-dimension hardness: $B = O(\sqrt{\log n})$, constant error.
- Fixed-d upper-bound exponents are *not* proved optimal.

Theorems 1.3, 5.4, 5.5; PDF pp. 3, 27. Adapted from Table 1, p. 5.

The algorithm makes exponential approximation local

1. **Trim the tail.** Retain scores close to the row maximum.
2. **Shift the row.** Relevant scores occupy a logarithmic interval.
3. **Approximate locally.** Use a low-degree polynomial for e^x.
4. **Aggregate moments.** Answer weighted geometric range queries.

Why all four steps matter

Scores can have huge absolute magnitude, yet the retained interval has width $O(\log(nB/\varepsilon))$.

Technical overview, PDF pp. 5–7; formal algorithm, Sec. 3, pp. 9–17.

A score gap controls the total discarded mass

Let $m_i = \max_j Q_i \cdot K_j$ and let η denote internal accuracy.

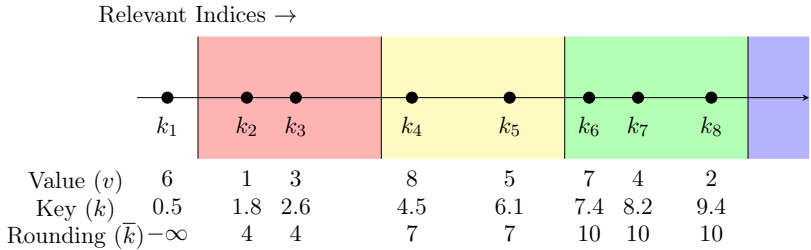
$$R_i = \{j : Q_i \cdot K_j \geq m_i - \log(n/\eta)\}.$$

$$j \notin R_i \implies \frac{p_{ij}}{p_{i,\max}} \leq \frac{\eta}{n} \implies \sum_{j \notin R_i} p_{ij} \leq \eta.$$

- Renormalizing retained weights changes an output entry by at most $3\eta B$.
- Values may be signed; the guarantee is additive.
- R_i can contain all keys: fast aggregation is still necessary.

Sec. 3.1, PDF pp. 10–11. The formal multidimensional algorithm uses a rounded row maximum.

In one dimension, relevance is an interval



- Positive query: retain a suffix in key order; negative query: reverse it.
- Figure 1 illustrates an initial rounding scheme, with polynomial accuracy cost.
- The final algorithm replaces fine bins with polynomial moments.

Original Figure 1, PDF p. 5; Secs. 1.1 and 3.1. Bin width 3 is illustrative, not a valid small-error setting.

After shifting, a logarithmic-degree polynomial suffices

$$c_i = m_i - \log(n/\eta), \qquad Q_i \cdot K_j - c_i \in [0, \log(n/\eta)] \quad (j \in R_i).$$

Choose P of degree $g = O(\log(n/\eta))$ satisfying

$$|P(x) - e^x| \le \eta e^x.$$

Normalized approximation

$$\widehat{O}_{i,t} = \frac{\sum_{j \in R_i} P(Q_i \cdot K_j - c_i) V_{j,t}}{\sum_{j \in R_i} P(Q_i \cdot K_j - c_i)}.$$

Shift invariance preserves the relevant softmax distribution.

Lemma 3.3 and Eq. (1), PDF p. 11; multidimensional interval, p. 14.

Sorted prefix moments give near-linear time for $d = 1$

$$\sum_{j \in R_i} P(q_i k_j - c_i) v_j = \sum_{\ell=0}^{g} m_\ell \sum_{b=0}^{\ell} \binom{\ell}{b} q_i^b (-c_i)^{\ell-b} \phi_v(i, b),$$

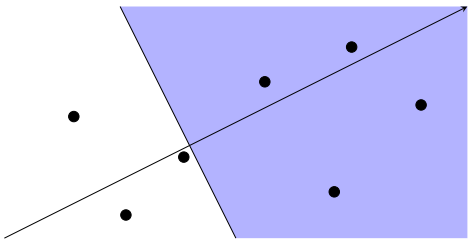
$$\phi_v(i, b) = \sum_{j \in R_i} k_j^b v_j.$$

- Sort keys once; store prefix sums of $k_j^b v_j$ and k_j^b.
- Binary search finds each query's interval endpoint.
- Include $b = 0$; a zero query returns the mean value.

$$\widetilde{O}(n \operatorname{polylog}(B/\varepsilon)) \text{ total time}$$

Lemmas 3.1, 3.4–3.5; Algorithm 1; PDF pp. 9–13.

In higher dimensions, relevance is a halfspace



$$R_i = \{K_j : Q_i \cdot K_j \geq c_i\}.$$

- Black points: keys.
- Arrow: query $(2, 1)$.
- Blue region: retained keys.
- Query directions differ; one sorted order no longer suffices.

Original Figure 2, PDF p. 7. Weights and coordinate scales are omitted in the original schematic.

Weighted range queries evaluate the polynomial

For a multi-index α, $|\alpha| \le g$, preprocess weights

$$w_{j,\alpha,t} = \left(\prod_{k=1}^{d} K_{j,k}^{\alpha_k} \right) V_{j,t}, \qquad \phi_t(i,\alpha) = \sum_{K_j \in R_i} w_{j,\alpha,t}.$$

- Multinomial expansion separates query factors from key moments.
- Build an extra set of moments without V for the denominator.
- Matoušek's structure: $O(n \log n)$ preprocessing per weight set; $\widetilde{O}(n^{1-1/d})$ per query.
- At most $g^{O(d)}$ moment types when d is fixed.

Theorem 3.7 and Lemmas 3.11–3.12, PDF pp. 13–16.

The error budget yields the main runtime bound

$$\underbrace{3\eta B}_{\text{discarded tail}} + \underbrace{4\eta B}_{\text{polynomial + normalization}} \le \varepsilon \quad \text{with} \quad \eta \le \frac{\varepsilon}{7B}.$$

$$\widetilde{O}\Big(n\, g^{O(d)} n^{1-1/d} \log(dB/\eta)\Big) = \widetilde{O}\Big(n^{2-1/d} \operatorname{polylog}(B/\varepsilon)\Big).$$

Fixed d	1	2	4	8
Exponent of n	1	3/2	7/4	15/8

The hidden $g^{O(d)}$ dependence matters for practical dimensions.

Eq. (1), p. 11; Lemmas 3.10, 3.13, pp. 14, 16. Exponent table is a reader calculation, not a benchmark.

Low rank reduces score geometry; output cost needs care

If $Q = U_Q V_Q^\top$ has rank r, then

$$Q' = U_Q \in \mathbb{R}^{n \times r}, \qquad K' = K V_Q \in \mathbb{R}^{n \times r}, \qquad Q' K'^\top = Q K^\top.$$

Theorem 1.4: reported bound

For $r = \min(\operatorname{rank} Q, \operatorname{rank} K) = O(1)$: $\widetilde{O}(nd + n^{2-1/r}\,\mathrm{polylog}(B/\varepsilon))$.

- Exact score reduction is valid; V still has d output columns.
- The printed factorization has incompatible matrix dimensions.
- The proof does not account for removing the per-column query cost.

Theorem 1.4 and proof, PDF pp. 16–17. Corrected score identity and qualifications: reader analysis.

The training result concerns one attention objective

$$H = A_3 Y, \quad F = \mathrm{softmax}_{\mathrm{row}}(A_1 X A_2^\top), \quad Z = FH, \quad C = Z - E,$$
$$L(X) = \tfrac{1}{2}\|C\|_F^2, \qquad \text{target: } \nabla_X L.$$

- Theorem 4.3 reports $O(d)$ attention calls plus $O(nd^2)$ work.
- Reported oracle accuracy: $\varepsilon/\Theta(ndB^3)$.
- The printed gradient identity is inconsistent with row softmax.

Scope

A single unit's gradient complexity does not establish end-to-end optimization cost or convergence for a full Transformer.

Definitions 4.1–4.2, Theorem 4.3, PDF pp. 17–19; full Transformer left open on p. 30.

A corrected identity preserves the few-call structure

Reader derivation: let $r_i = \langle C_i, Z_i \rangle$ and compute

$$U = FA_2, \qquad T_p = F\,\mathrm{diag}(H_{*,p})A_2.$$

$$\nabla_X L = A_1^\top \left[\sum_{p=1}^{d} \mathrm{diag}(C_{*,p})\, T_p - \mathrm{diag}(r)U \right].$$

- Z, U, and the d matrices T_p: $d+2$ attention calls.
- Remaining row operations and multiplication cost $O(nd^2)$.
- Intermediate entry bounds grow; accuracy requires fresh bookkeeping.

Reader derivation from Definition 4.1. The printed second term on pp. 18–19 differs; numerical check in backup.

A sentinel key exposes an unnormalized row sum

Let $S_i = \sum_j e^{Q_i \cdot K_j}$. Add a key with score $\ln c_i$.

Original keys have value 1; the sentinel has value 0.

$$o_i = \frac{S_i}{c_i + S_i}, \qquad o_i > \tfrac{1}{2} \iff S_i > c_i.$$

- A small additive error distinguishes multiplicatively separated thresholds.
- Binary search on a geometric grid recovers a relative row-sum estimate.
- The added coordinate $\ln c_i$ may exceed the original entry bound.

Lemma 5.1 and Claim 5.2, PDF pp. 25–26. Entry-bound qualification: reader verification.

An exponential sum can reveal an integer maximum

For integer inputs, let $M_i = \max_j a_i \cdot b_j$ and choose constant C_0:

$$S_i = \sum_j n^{C_0(a_i \cdot b_j)}, \qquad n^{C_0 M_i} \le S_i \le n^{C_0 M_i + 1}.$$

Taking a logarithm and rounding recovers M_i, even from a constant-relative-error estimate of S_i.

Conditional hardness with very slowly growing dimension

For every $\delta > 0$, there is $C = C(\delta)$ such that $d = 2^{C \log^* n}$ requires $n^{2-\delta}$ time under SETH.

Here $B = \text{poly}(n)$ and $\varepsilon = \Theta(1)$; the enlarged sentinel bound remains polynomial.

Theorem 2.3, p. 9; Lemma 5.3 and Theorem 5.4, PDF pp. 26–27.

Large dimensions inherit matrix-multiplication hardness

Boolean vectors with scores $\ln(k)(a_i \cdot b_j)$ give

$$S_i = \sum_j k^{a_i \cdot b_j}, \qquad k = \frac{1}{10n}.$$

No orthogonal pair	An orthogonal pair exists
Every row has $S_i \le nk = 0.1$.	Some row has $S_i \ge 1$.

A fixed sentinel with $c = 1/2$ distinguishes the cases at constant accuracy.

- $B = O(\sqrt{\log n})$; input dimensions become $n + 1, d + 1$.
- Theorem 5.5: $T_{\mathrm{MUL}}(n, d, n)^{1-o(1)}$ under generalized high-dimensional OV.

Theorem 5.5 and Eq. (5), PDF pp. 27–28. Fixed-threshold sentinel: reader simplification.

The constant-entry claim depends on delicate parameters

The intended encoding is an integer row sum:

$$S_i = \sum_j 2^{a_i \cdot b_j}.$$

- Nonorthogonal pairs contribute even integers; orthogonal pairs contribute 1.
- Independent half-sampling detects an odd count with probability $1/2$.
- Inverse-polynomial accuracy allows exact rounding when $d = O(\log n)$.

Version-1 proof qualification

Both Q and K are printed with scale $\log 2$, producing $(\log 2)^2$ scores. The claimed $B = \log 2$ bound also conflicts with the sentinel construction.

Theorem 5.7 and Lemma 5.8, PDF pp. 28–29. Exact constant-bound statement not verified in this reading.

The evidence is asymptotic and conditional

- **No empirical evaluation:** no model training, hardware benchmark, or measured speedup.
- **Fixed dimension:** $g^{O(d)}$ can dominate at realistic head sizes.
- **Worst-case guarantees:** entrywise accuracy and input assumptions define the lower bounds.
- **Unresolved interval:** $1 \ll d \ll 2^{\Theta(\log^* n)}$.
- **Large-d optimality:** relative to fast matrix multiplication; for $d = n$, the scale is n^ω, not n^3.

Table 1, PDF p. 5; Secs. 3 and 5; Conclusion, p. 30. Practical interpretation: reader analysis.

The lasting idea is query-specific local approximation

1. **Algorithm:** tail bounds create short score windows; geometry aggregates polynomial moments.
2. **Complexity:** fixed dimension permits truly subquadratic attention with polylogarithmic magnitude dependence.
3. **Hardness:** exponential weighting connects attention to integer Max-IP and OV.
4. **Reading discipline:** preserve the hypotheses and verify extensions before using their strongest claims.

Discussion

Can the local-window idea produce a practical algorithm with a useful, explicitly stated accuracy guarantee?

Main algorithm: Theorem 1.3; reductions: Sec. 5. Discussion question: reader analysis.

Sources and reading coverage

Primary paper

Gupta, Huang, Saha, Xu, Ye. *Subquadratic Algorithms and Hardness for Attention with Any Temperature*. arXiv:2505.14840v1.

- Read all substantive sections, PDF pp. 1–30, using author TeX and PDF.
- Figures 1–2 are original PDF crops; Table 1 was checked.
- Core tools cited by the paper: Matoušek (1992), Aggarwal–Alman (2022), Chen (2018).
- External normalization check only: Alman–Song, arXiv:2302.13214v2.

Reader analysis and numerical checks are explicitly labeled. No independent implementation or benchmark is claimed.

Backup: normalization contributes the second error term

For retained probabilities,

$$\frac{1-\eta}{1+\eta} p_{ij}^{(r)} \le \widehat{p}_{ij} \le \frac{1+\eta}{1-\eta} p_{ij}^{(r)}.$$

For $\eta < 0.1$, this implies

$$|\widehat{O}_{i,t} - O_{i,t}^{(r)}| \le B \sum_{j \in R_i} |\widehat{p}_{ij} - p_{ij}^{(r)}| \le 4\eta B.$$

- Add the tail bound $3\eta B$ to obtain $7\eta B$.
- Signed values require an additive-output argument.
- With $\eta \asymp \varepsilon/B$, $g = O(\log(nB/\varepsilon))$.

Sec. 3.1, Eq. (1), PDF p. 11; Lemma 3.10, p. 14.

Backup: locate a maximum without enumerating scores

Find the largest integer s_i such that the halfspace

$$\{x : Q_i \cdot x \geq s_i \log(1+\eta)\}$$

contains a key. Since $|Q_i \cdot K_j| \leq dB^2$, binary search needs $O(\log(dB/\eta))$ count queries.

$$c_i = s_i \log(1+\eta) - \log(n/\eta).$$

- Relevant shifted scores lie in $[0, \log(n/\eta) + \log(1+\eta)]$.
- Reuse the same halfspace for every weighted moment.
- A zero query can return the value mean directly.

Definition 3.8 and Lemma 3.12, PDF pp. 13, 15–16. Zero-query handling: reader implementation note.

Backup: what is missing from the low-rank proof?

- Printed product: $V_Q^\top U_K V_K^\top$. The first two factors have dimensions $r_Q \times d$ and $n \times r_K$.
- The proof treats $\max(r_Q, r_K)$ as constant, while the theorem assumes only their minimum is constant.
- With $V \in \mathbb{R}^{n \times d}$, the moment construction still has d output channels.

A conservative direct extension

One-sided score reduction plus separate output channels gives a bound of the form $\widetilde{O}(nd + d\, n^{2-1/r}\, \mathrm{polylog}(B/\varepsilon))$.

A stronger batched result would need an additional argument.

PDF pp. 16–17. Dimension check and conservative bound: reader analysis, not a lower bound or an author correction.

Backup: the printed gradient fails a small algebra check

$n = 2, d = 1$, $A_1 = (1, 2)^\top$, $A_2 = (-1, 1)^\top$,
$H = (1, 3)^\top$, $E = (0.2, -0.1)^\top$, and $X = 0.4$.

Calculation	Derivative
Centered finite difference ($h = 10^{-5}$)	4.9557469640
Correct row-softmax identity	4.9557469642
Printed identity, PDF pp. 18–19	−0.4520133397

Correct subtraction: $\operatorname{diag}((F \circ CH^\top)\mathbf{1})F$.

- Finite differences also agree at $h = 10^{-4}$ and 10^{-6}.
- This checks an identity, not the impossibility of a gradient algorithm.

Reader's reproducible example: build/verify_math.py and build/math-check.json. Not a paper experiment.

Backup: parameter changes matter in hardness reductions

- **Generic sentinel:** $\ln c_i$ can have magnitude $O(\log n + dB^2)$; the original bound B is not automatically preserved.
- **Polynomial-entry Max-IP:** this enlargement stays polynomial, preserving the relevant parameter regime.
- **High-dimensional OV:** a fixed threshold suffices; $S \leq 0.1$ gives $o \leq 1/6$, while $S \geq 1$ gives $o \geq 2/3$.
- **Constant-entry parity:** using $\sqrt{\log 2}$ on both sides fixes integer scores but exceeds the printed $B = \log 2$ threshold.

Lemma 5.1, Theorems 5.5 and 5.7, PDF pp. 25–29. Parameter audit and fixed-threshold repair: reader analysis.