# ROBUST DECISION CAPACITY OF MULTI-HEAD ATTENTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We establish head-count limits on stable threshold decisions in a single attention layer, through its response to changing one source value along an affine query path while holding keys and scores fixed during the intervention. We quantify stable decision complexity by robust alternating capacity: the maximal number of transitions between alternating threshold labels, each certified under every independent head gain in $[1-\delta, 1+\delta]$. Two heads can approximate every smooth scalar response in $C^r$ on a compact interval, yet robust decision capacity remains limited by head count. For $H \geq 2$ and $0 < \delta \leq 1/8$, we prove that unrestricted-score capacity is $\Theta(H\delta^{-1}\log(1/\delta))$. The upper bound holds for every output norm and is independent of width, context length, and output-coefficient magnitude; finite scalar attention constructions attain the same order. With source-relative score slopes bounded by $B$ on a path of length $L$, we derive matching Euclidean capacity laws in joint head–score regimes, including a transition from factorially constrained growth to a square-root regime governed by $H$, $BL$, and $\log(1/\delta)$. Fixed-head high-score laws reach the unrestricted capacity order, while a fixed non-Euclidean output norm changes the score-dependent order. These results show that head count controls stable decision complexity even after smooth approximation has saturated.

## 1 INTRODUCTION

*Head count controls the gain-stable decision capacity of a localized source response, even with arbitrary finite width and context length.* We prove this through a universal ceiling independent of width, context length, and output-coefficient magnitude, together with finite constructions attaining its order. At a fixed gain tolerance, maximal robust alternation grows linearly with head count even though two heads already approximate every smooth scalar response in this interface. With bounded score variation, matching bounds further quantify the head–score tradeoff.

Head-count hierarchies are known for discrete attention-only computation even with unrestricted embedding dimension and precision (Rajaraman et al., 2026); approximation rates can also depend on per-head resources (Amsel et al., 2025; Yu et al., 2026). We study head count in a continuous interface where ordinary approximation has already saturated: the localized response of one source token along an affine query path, allowing arbitrary finite width and context length. Two heads are $C^r$-dense in its smooth scalar responses. We ask how head count limits *stable decision complexity* in this interface.

To make this question precise, change one source value payload while holding queries, keys, and scores fixed, and observe a linear readout. The response is a sum of source softmax weights multiplied by fixed head maps. Head-wise scalar gates are used to probe individual contributions (Michel et al., 2019; Voita et al., 2019); here we restrict each gain to $[1-\delta, 1+\delta]$. We count transitions between alternating threshold labels along the path only when each selected label remains correct for every gain vector. The supremum over all finite attention realizations, thresholds, and ordered witnesses is the robust alternating capacity. Oscillation counts along lines are a structural measure of decision complexity (Telgarsky, 2016); the gain box adds a head-decomposition stability requirement while leaving attention patterns unchanged.

Our two-head density construction uses opposing readout coefficients whose magnitudes grow as approximation is refined. Robust alternating capacity adds a uniform gain-margin requirement to



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the threshold labels. We establish its head-count dependence through a universal upper bound and a matching finite construction: for $H \geq 2$ and $0 < \delta \leq 1/8$, unrestricted-score capacity is

$$\Theta\big(H\delta^{-1}\log(1/\delta)\big)\,.$$

Each robust transition spends part of a logarithmic variation budget, while the unimodality of each source weight limits the total available per head. The upper bound is independent of coefficient magnitude, width, and context length; finite constructions match it in standard causal scaled-dot-product attention. Together, these bounds characterize the attainable capacity order across the entire class.

Unrestricted scores can vary arbitrarily fast along the query path. To measure the cost of stable decisions, let $B$ bound every source-relative affine score slope on a path of length $L$, and put

$$D = \log(1/\delta), \qquad x = BL.$$

The product $x$ is unchanged by rescaling the path coordinate. For $D \geq 32$, $2 \leq H \leq D$, and $1 \leq x \leq HD$, we prove the uniform Euclidean capacity law

$$\mathcal{C}^{\mathrm{Euc}}_{B,L}(H,\delta) = \begin{cases} \Theta\left(\dfrac{D}{\log(e + D/(Hx))}\right), & Hx \leq D, \\ \Theta(\sqrt{HxD}), & Hx > D. \end{cases}$$

The constants are independent of $H$. The crossover at $Hx = D$ separates a factorially constrained low-score regime from a square-root regime. For example, at $x = D$, increasing head count from two to $H \asymp D$ raises capacity from $\Theta(D)$ to $\Theta(D^{3/2})$.

For the square-root upper bound, clearing softmax denominators produces a positive spectral envelope whose log curvature is Fisher information. Exponential tilting and zero counting charge transitions against its integrated information budget of $O(HB)$, independently of the number of exponential terms (Lemma S1). For the lower bound, a finite Gaussian lattice supplies repeated excursions, while a $k$-head rational function creates $2(k-1)$ robust transitions per excursion.

For high score budgets and fixed $H \geq 2$, reparameterize the score product by $G = \log(1+BL/(8D))$. Two scalar heads attain the matching Euclidean laws

$$\mathcal{C}^{\mathrm{Euc}}_{B,L}(H,\delta) = \begin{cases} \Theta_H(DG), & 1 \leq G \leq D, \\ \Theta_H(G[1+\log(D/(\delta G))])\,, & D \leq G \leq D/\delta, \end{cases}$$

for $D \geq 32$. A Mellin pair gives the first branch. In the second, a square-envelope correction encodes the threshold baseline inside two positive denominators; local reciprocal approximation gives the matching upper bound. At $G = D/\delta$ the rate reaches the unrestricted order.

At the low-score boundary, $x \leq \log((1+\delta)/(1-\delta))$ forbids every transition. The score-dependent rates use Euclidean output geometry: a fixed non-Euclidean norm yields a different order even with two heads.

## 2 Localized Source Dependence and Robust Capacity

Consider one causal attention layer with $H$ heads. A scalar parameter $z \in I \subset \mathbb{R}$ controls the final query through an affine path. Each score is therefore affine,

$$\ell_{hj}(z) = a_{hj}z + b_{hj}. \tag{1}$$

Choose one source token $s$ and intervene on a designated value payload $v$ while holding keys, queries, and all logits fixed. Background values cancel from the intervention difference, so for vector payloads and outputs

$$F(z, v+\Delta v) - F(z,v) = T(z)\Delta v, \qquad T(z) = \sum_{h=1}^{H} w_h(z)B_h, \tag{2}$$

where $w_h$ is the source softmax weight and $B_h$ is the effective linear map from the source payload through head $h$ to the observed output. For every intervention radius $\rho > 0$, the normalized finite-radius effect is exactly $\|T(z)\|_{\mathrm{op}}$.

The source weight has the form

$$\log w_h(z) = \ell_{hs}(z) - \log \sum_j e^{\ell_{hj}(z)}. \tag{3}$$

The right-hand side is concave; hence $w_h$ is nonnegative and unimodal on $I$. These two properties suffice for the universal upper bound below.

Fix a threshold $\varepsilon > 0$ and perturb the head contributions by gains $t_h \in [1-\delta, 1+\delta]$, where $0 < \delta < 1$. Define

$$A_t(z) = \left\| \sum_{h=1}^{H} t_h w_h(z) B_h \right\|_{\mathrm{op}}. \tag{4}$$

**Definition 1** (Robust witnesses)**.** *A point $z$ is* robustly active *if $A_t(z) > \varepsilon$ for every admissible $t$, and* robustly inactive *if $A_t(z) \le \varepsilon$ for every admissible $t$. An ordered witness sequence $z_0 < \cdots < z_K$ is alternating when consecutive labels differ. We count its $K$ transitions.*

Let $\mathfrak{A}_H$ denote one-layer causal scaled-dot-product attention maps with exactly $H$ heads, arbitrary finite context and width, finite real projection matrices, the interface above, and a fixed linear observed readout. Gains are inserted immediately before the output projection. We write $\mathcal{C}_{\mathrm{attn}}(H, \delta)$ for the supremum of $K$ over $\mathfrak{A}_H$, all fixed positive thresholds, and all alternating witness sequences.

The definition certifies decisions at selected ordered points while allowing threshold boundaries to move and shallow regions to appear between witnesses. Witness locations and the threshold are optimized together with the attention map. Common rescaling of the output and threshold leaves every witness inequality unchanged.

## 3 Smooth Expressivity Saturates at Two Heads

**Theorem 2** (Two-head smooth density)**.** *Let $I$ be a compact interval, let $r \ge 0$ be an integer, and let $f$ be $C^r$ on a neighborhood of $I$. For every $\gamma > 0$, a finite causal attention layer with two heads has a scalar localized-source response $g$ such that $\|g - f\|_{C^r(I)} < \gamma$. Its affine query chart is fixed independently of $f$.*

The construction is a difference of two reciprocal positive exponential sums. Polynomial approximation after $y = e^z$ first gives $R(z) = \sum_{k=0}^{n} r_k e^{kz}$ close to $f$ in $C^r(I)$. Choose positive exponential sums $A_+, A_-$ with $A_+ - A_- = R$, and set $P_\lambda = 2 + \lambda A_-$ and $Q_\lambda = 2 + \lambda A_+$. Then

$$g_\lambda(z) = \frac{4}{\lambda}\left(\frac{1}{P_\lambda(z)} - \frac{1}{Q_\lambda(z)}\right) = \frac{4R(z)}{P_\lambda(z) Q_\lambda(z)} \longrightarrow R(z) \quad \text{in } C^r(I). \tag{5}$$

Each reciprocal is a source softmax weight; the factor 2 accounts for the source and query tokens. Appendix A provides the finite projection matrices and the approximation details.

One scalar head has fixed sign, so it cannot approximate a sign-changing target. Two heads overcome this sign obstruction. The coefficients $\pm 4/\lambda$ in (5) grow as the approximation is refined. The capacity defined in Section 2 adds independent-gain stability to the threshold labels, distinguishing smooth approximation from robust decision complexity.

## 4 Sharp Capacity with Unrestricted Scores

**Theorem 3** (Unrestricted robust capacity)**.** *For $H \ge 2$ and $0 < \delta \le 1/8$, put $D = \log(1/\delta)$ and $Q_\delta = \log((1-\delta)/\delta)/\log((1+\delta)/(1-\delta))$. Then*

$$2\lfloor H/2 \rfloor (\lceil Q_\delta \rceil - 1) \le \mathcal{C}_{\mathrm{attn}}(H, \delta) \le 2H + \frac{H}{\delta} \inf_{r>0} (1+r) \log\left(1 + \frac{1}{\delta r}\right). \tag{6}$$

*The upper bound holds for every norm on a finite-dimensional output space. The lower bound uses scalar outputs and finite standard attention. Thus the capacity is $\Theta(H\delta^{-1}\log(1/\delta))$, with absolute constants. One head has capacity at most two.*

**The logarithmic variation budget.** The upper bound uses only nonnegativity and unimodality of each source weight. At an inactive witness, a norming functional for head $h$ and the gain-box support function give $\|B_h\| w_h \le \varepsilon/(2\delta)$. Unimodality lets us remove at most $2H$ edges to enforce this cap at every remaining endpoint. Define

$$\tau > 0, \qquad \Phi(a) = \log(1 + a/\tau).$$

For a retained edge with active endpoint $p$ and inactive endpoint $q$, convex separation gives a dual-unit functional $\ell$ that separates the entire active gain-response set from the threshold ball. Put $b_h = |\ell(B_h)|$. Comparing the two gain-box extrema gives

$$\sum_h b_h |w_h(p) - w_h(q)| > \delta \sum_h b_h (w_h(p) + w_h(q)).$$

Write $S = \sum_h b_h (w_h(p) + w_h(q)) > \varepsilon$. The logarithmic inequality $\log(x/y) \ge 2(x-y)/(x+y)$ for $x \ge y > 0$ then yields the edge charge

$$\sum_h |\Phi(\|B_h\| w_h(p)) - \Phi(\|B_h\| w_h(q))| > \frac{2\delta S}{2\tau + S} > \frac{2\delta}{1 + 2\tau/\varepsilon}. \tag{7}$$

The separating functional can change from edge to edge, but the potential in (7) is fixed across the whole sequence. This removes the dependence on the directions and magnitudes of the head maps. Each capped potential has total variation at most $2\log(1 + \varepsilon/(2\delta\tau))$. Summation and optimization over $r = 2\tau/\varepsilon$ give the upper bound, without context-length or output-dimension dependence. Its optimized logarithmic factor is $D + \log D + O(1)$ as $\delta \downarrow 0$.

**Matching the budget with decreasing head amplitudes.** For the lower bound, choose $a > \delta$ sufficiently close to $\delta$, set $r = (1-a)/(1+a)$, and

$$T_i = \frac{2\varepsilon}{a + \delta} r^i, \quad g_i = \varepsilon + (-1)^i a T_i, \quad u_i = \frac{T_i + g_i}{2}, \quad v_i = \frac{T_i - g_i}{2}.$$

As long as $T_i > \varepsilon/(1-a)$, both $u_i, v_i$ are positive and nonincreasing. Their difference crosses the threshold repeatedly even though each sequence decreases. More precisely, all independently perturbed differences lie in $[g_i - \delta T_i, g_i + \delta T_i]$, inside $(\varepsilon, \infty)$ for even $i$ and inside $(-\varepsilon, \varepsilon)$ for odd $i$. The decay ratio makes $T_i - T_{i+1} = |g_i - g_{i+1}|$, using the full monotonicity allowance.

Symmetric positive exponential sums interpolate these reciprocal amplitudes on both sides of their common peak. Reflection doubles the transition count with the same two heads. Taking a sufficiently large finite score scale preserves every strict gain margin. Localizing $\lfloor H/2 \rfloor$ pairs then supplies the factor $H$. Appendix B gives the interpolation and explicit constants.

For scalar head coefficients, the leading constant is also sharp:

**Corollary 4** (Two scalar heads)**.** *As $\delta \downarrow 0$,*

$$\mathcal{C}^{\mathrm{sc}}_{\mathrm{attn}}(2, \delta) = \delta^{-1} \log(1/\delta) + O(1), \tag{8}$$

*where $\mathcal{C}^{\mathrm{sc}}_{\mathrm{attn}}$ restricts the head coefficients to scalars.*

For $0 < \delta \le 1/8$, the finite bracket is $2\lceil Q_\delta \rceil - 2 \le \mathcal{C}^{\mathrm{sc}}_{\mathrm{attn}}(2, \delta) \le 2\lceil Q_\delta \rceil + 6$; its proof is in Appendix B.

## 5 A Joint Law under Bounded Score Variation

Suppose all source-relative slopes obey $|a_{hj} - a_{hs}| \le B$ on an interval of length $L$. Write $x = BL$ and $D = \log(1/\delta)$. Let $\mathcal{C}^{\mathrm{Euc}}_{B,L}(H, \delta)$ denote the capacity with scalar source payload and Euclidean output norm; the lower bounds use scalar outputs. Intercepts, output coefficients, finite width, and finite context length remain unrestricted. The quantity $x$ is invariant under rescaling the query coordinate.

**Theorem 5** (Joint score–head capacity)**.** *For $0 < \delta \le 1/8$, $H \ge 2$, and $x > 0$, put $X = 4Hx$, $\rho = \delta^3(2+\delta)/[(1+\delta)^2(H+\delta^2)]$, and $M_X = \max\{j \in \mathbb{Z}_{\ge 0} : \rho(j+1)! \le e^X X^j\}$. Then*

$$\mathcal{C}^{\mathrm{Euc}}_{B,L}(H, \delta) \le \min\left\{1 + M_X, \left\lfloor \min_{n \in \mathbb{N}} \left[ n - 1 + \frac{n \log(1/\rho) + 12X/n}{\log 2} \right] \right\rfloor \right\}. \tag{9}$$

*In particular, this is $O(U+\sqrt{HxU})$, $U = D+\log(H+1)$. For $D \geq 32$, uniformly over $2 \leq H \leq D$ and $1 \leq x \leq HD$, the matching capacity law is*

$$\mathcal{C}^{\mathrm{Euc}}_{B,L}(H,\delta) = \begin{cases} \Theta\left(\dfrac{D}{\log(e + D/(Hx))}\right), & Hx \leq D, \\ \Theta\left(\sqrt{HxD}\right), & Hx > D. \end{cases} \tag{10}$$

*All implicit constants are independent of $H, x, \delta$.*

The low-score endpoint has a sharp threshold. Put $\tau_\delta = \log((1+\delta)/(1-\delta))$. For every output norm and $H \geq 2$, every nonempty alternating sequence satisfies

$$K < x/\tau_\delta, \qquad \mathcal{C}^{\mathrm{Euc}}_{B,L}(H,\delta) = 0 \iff x \leq \tau_\delta. \tag{11}$$

The remaining low-score interval has a matching law for every head count: for $D \geq 64$, uniformly over $H \geq 2$ and $\tau_\delta < x \leq 1$,

$$\mathcal{C}^{\mathrm{Euc}}_{B,L}(H,\delta) = \Theta\left(\frac{D}{\log(e + (1 + D/H)/x)}\right). \tag{12}$$

Appendix G proves both statements. Two transitions begin exactly at $x > 2\tau_\delta \asymp \delta$ for $H \geq 2$, versus $x > 2\,\mathrm{arcosh}(e^{\tau_\delta}) \asymp \sqrt{\delta}$ for one head.

The two rates meet at $Hx = D$, where capacity is $\Theta(D)$. At $x = D$, increasing head count from $H = 2$ to $H = \lfloor D \rfloor$ raises the capacity rate from $D$ to $D^{3/2}$.

**An information budget for robust oscillations.** Write the response as $\sum_h b_h/Q_h$, where each $Q_h$ is a positive exponential sum with exponents in $[-B, B]$. Clear denominators using

$$Q = \prod_h Q_h, \qquad N = \sum_h b_h Q/Q_h, \qquad F = \|N\|_2^2 - \varepsilon^2 Q^2.$$

The signs of $F$ alternate at the robust witnesses. A positive coefficient envelope is

$$A = \varepsilon^2 Q^2 + \left(\sum_h \|b_h\|_2 Q/Q_h\right)^2.$$

It has exponent diameter at most $4HB$. Averaging the squared response over the corners of the gain box bounds the envelope at inactive witnesses, giving the relative negative margin $\rho$ above. To expose the envelope's geometry, write $A(z) = \sum_j a_j e^{\lambda_j z}$ with $a_j > 0$. It induces the spectral exponential family $\pi_z(j) = a_j e^{\lambda_j z}/A(z)$. Its Fisher information satisfies

$$\mathcal{I}_A(z) := \sum_j \pi_z(j)(\partial_z \log \pi_z(j))^2 = \mathrm{Var}_{\pi_z}(\lambda) = (\log A)''(z), \qquad \int_{\mathbb{R}} \mathcal{I}_A(z)\,dz \leq 4HB.$$

These identities (Wainwright & Jordan, 2008) give the information budget. On each of $M$ equal subintervals, tilt $F$ by the tangent slope of $\log A$. Its local growth is a KL divergence; a disk zero estimate and bounded overlap give

$$K \leq M - 1 + \frac{M\log(1/\rho) + 48Hx/M}{\log 2}.$$

Optimizing in $M$ gives (9), without a term-count dependence. Appendix C gives the local estimate.

Below the crossover, a divided difference provides a stronger constraint. Starting at an inactive witness and shifting all exponents to $[0, 4HB]$ makes every witness value have magnitude at least $\rho A(z_0)$. The alternating signs make all divided-difference summands agree in sign. For $m \geq K - 1$ remaining transitions and $X = 4Hx$, this gives

$$\rho(m+1)m! \leq e^X X^m.$$

Maximizing the feasible integer $m$ gives $K \leq 1 + M_X$; Stirling inversion when $Hx \leq D$ yields the first line of (10). Appendix C proves both upper bounds with constants uniform in the head count.

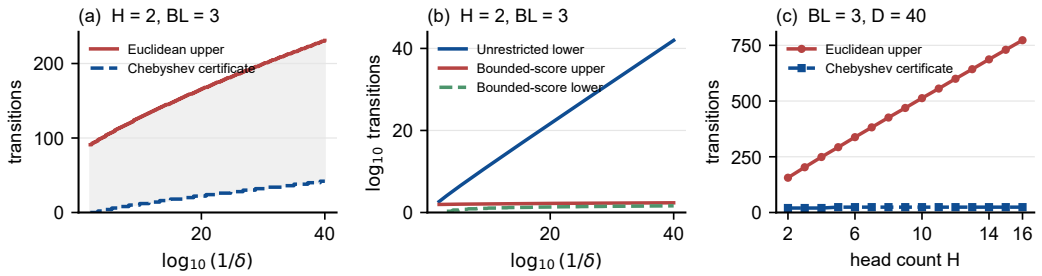



Figure 1: Finite-parameter consequences of the capacity laws. **(a,b)** At $H = 2$, $BL = 3$, compare the Euclidean upper bound (9), the exact Chebyshev lower certificate $K_C$ (Appendix D), and the unrestricted lower bound (6). **(c)** Varying $H$ at $BL = 3$, $D = \log(1/\delta) = 40$.

**How the construction attains the rate.** Choose exponential sums $P, A$ whose coefficients satisfy $|P_j| \le A_j$ and for which $r = P/A$ repeatedly traverses $[-a, a]$. Pick distinct Chebyshev nodes $c_j \in (0, 1/2)$. The proper rational function

$$\frac{T_{k-1}(r/a)}{\prod_{j=1}^{k}(1 + c_j r)} = \sum_{j=1}^{k} \frac{\beta_j}{1 + c_j r} \tag{13}$$

becomes an attention response after division by $A$: head $j$ uses denominator $Q_j = A + c_j P$ and readout $\beta_j$. Positivity holds coefficient by coefficient. The numerator alternates between extrema and zeros $2(k-1)$ times per traversal, giving alternating threshold labels. Define $P_k(a) = \prod_j (1 + c_j a)$ and $S_k(a) = \sum_j |\beta_j|/(1 - c_j)$, and fix $0 < \gamma < 1$. With $R = A_{\max}/A_{\min}$, the finite sufficient condition

$$2\delta R P_k(a) S_k(a) \le \gamma \tag{14}$$

leaves normalized slack at least $1 - \gamma$ and guarantees $2m(k-1)$ transitions for $m$ traversals. Appendix D gives the residues explicitly.

For the Gaussian construction, take even $m$, $0 < \eta \le 1$, $p = \lceil \pi/\eta \rceil$, and $s = mz/L$, with

$$A(s) = \sum_{j=-p}^{m+p} e^{\eta(2js - j^2 - ms)}, \qquad P(s) = \sum_{j=-p}^{m+p} (-1)^j e^{\eta(2js - j^2 - ms)}.$$

Define $\Theta_\eta = \sum_{j\in\mathbb{Z}} e^{-\eta j^2}$, $C_\eta = \sum_{j\in\mathbb{Z}} (-1)^j e^{-\eta j^2}$, and $E_{\eta,p} = 2\sum_{j>p} e^{-\eta j^2}$. At successive integers, $P/A$ alternates beyond $\pm a_{\eta,p}$, $a_{\eta,p} = (C_\eta - E_{\eta,p})/\Theta_\eta > 0$. The finite conditions

$$\eta m(m + 2p) \le x, \qquad 2\delta \Theta_\eta e^{\eta(m^2+1)/4} P_k(a_{\eta,p}) S_k(a_{\eta,p}) < 1$$

guarantee $K \ge 2m(k-1)$ with $k \le H$. The largest score-feasible even integer is $m = 2\lfloor(\sqrt{p^2 + x/\eta} - p)/2\rfloor$; smaller even integers remain eligible if the gain condition fails. Maximizing this integer count over feasible $k, m, \eta$ gives the finite lower bound. The choices justified in Appendix D yield $\Omega(\sqrt{HxD})$ for $D/H \le x \le D$; localization retains $2Jm(k-1)$ transitions from $J$ copies. Below $Hx = D$, a transformed Chebyshev sum realizes the factorial branch. Appendix D gives the uniform parameters and finite realizations. Figure 1 evaluates these finite capacity certificates.

## 6 A Matching High-Score Law with Two Heads

We next fix two heads and increase the score budget. A square-envelope construction combines a threshold baseline with repeated oscillations and matches the fixed-head upper bound.

A two-head construction also supplies a lower bridge to that range. For $D \ge 32$, $1 \le G \le D$, $B = 8D$, and $L = e^G - 1$, finite scalar attention realizes

$$\mathcal{C}^{\mathrm{Euc}}_{B,L}(2, \delta) \ge 2\lfloor k_\dagger/\pi \rfloor = \Omega(DG). \tag{15}$$

Here $k_\dagger > 0$ solves $\log(1 + k_\dagger) + \pi k_\dagger/(2G) = D - \log(4e + 2)$. Appendix H proves the finite construction. For each fixed $H_0 \geq 2$, a bounded-envelope upper bound matches it: uniformly over $2 \leq H \leq H_0$ at the same score product,

$$\mathcal{C}^{\mathrm{Euc}}_{B,L}(H, \delta) = \Theta_{H_0}(DG), \qquad 1 \leq G \leq D. \tag{16}$$

The upper proof is in Appendix I.

**Theorem 6** (Fixed-head high-score law)**.** *Let $D = \log(1/\delta) \geq 20$ and $D \leq G \leq D/\delta$. Put $a = 2D/3 - \log 128$ and let $k_* > 0$ solve $\log(1+k_*/a)+\pi k_*/(2G) = D-\log 16$. At $BL = 8D(e^G-1)$, two scalar heads in finite standard attention attain*

$$K \geq 2\lfloor k_*/(2\pi) \rfloor = \Omega\left( G \left[ 1 + \log \frac{D}{\delta G} \right] \right). \tag{17}$$

*The construction uses $\lceil [\log(8D) + 3DG/a]/h \rceil + 3$ tokens, where Appendix E specifies $h$ by the finite error budget; this is $O(G/\sqrt{\delta})$. For every fixed $H_0 \geq 2$, uniformly over $2 \leq H \leq H_0$ at the same score product,*

$$\mathcal{C}^{\mathrm{Euc}}_{B,L}(H, \delta) = \Theta_{H_0}\left( G \left[ 1 + \log \frac{D}{\delta G} \right] \right). \tag{18}$$

At $G = D$, the theorem gives $\Omega(D^2)$ transitions at score budget $x = 8D(e^D - 1)$. Since $\log x = D + \log(8D) + \log(1 - e^{-D})$, this rules out a uniform additive law $O(D + \log x)$.

**Putting the baseline into the denominators.** Let $A(z) > 0$ be a common envelope and introduce two ideal denominators $Q_\pm = A \pm F$. Opposite readout coefficients give

$$R(z) = \frac{\beta}{A - F} - \frac{\beta}{A + F} = \frac{2\beta F}{A^2 - F^2}.$$

The term $A^2/(2\beta)$ inside $F$ creates a baseline of one; the remaining oscillation crosses the threshold while the reciprocal correction $1/(1 - (F/A)^2)$ stays within the gain margin.

Set

$$A(z) = \Gamma(\alpha) z^{-\alpha}, \qquad F(z) = \frac{A(z)^2}{2\beta} + \delta A(z)\{8\cos(\omega \log z) - 4\}, \qquad 1 \leq z \leq e^G.$$

The Mellin–Laplace identity represents both the oscillation and $A^2$ by positive denominator densities after a phase choice; the required frequency satisfies $|\Gamma(\alpha + i\omega)|/\Gamma(\alpha) = 16\delta$.

**The gain margin absorbs the reciprocal error.** Write $t = 2\beta/A$ and $h = F/A$. The nominal response and total absolute head amplitude are

$$R = \frac{1 + \delta\{8\cos(\omega \log z) - 4\}t}{1 - h^2}, \qquad T = \frac{t}{1 - h^2}. \tag{19}$$

Independent gains give exactly $[R - \delta T, R + \delta T]$. Take $t_0 = 2\delta^{-1/3}$, $a = 2D/3 - \log 128$, $\alpha = a/G$, and $\beta = t_0\Gamma(\alpha)/2$. Throughout the path,

$$2\delta^{-1/3} \leq t \leq (64\delta)^{-1}, \qquad |h| \leq \delta^{1/3}, \qquad h^2 \leq \delta t/2.$$

At a cosine zero, the entire gain interval lies in $(0, 1)$ because

$$1 - (R + \delta T) = \frac{3\delta t - h^2}{1 - h^2} > 0, \qquad R - \delta T = \frac{1 - 5\delta t}{1 - h^2} > 0.$$

At a positive maximum, $R - \delta T - 1 = (3\delta t + h^2)/(1 - h^2) > 0$. Alternating these two kinds of witnesses guarantees $2\lfloor \omega G/(2\pi) \rfloor$ transitions. The gamma-product estimate gives

$$\omega G \geq cG \left[ 1 + \log\left( \frac{D}{\delta G} \right) \right],$$

which is the claimed rate.

Appendix E gives the positive finite realization and frequency bound; Appendix F proves the matching upper bound.

## 7 Output Geometry Changes the Score-Dependent Order

The unrestricted upper bound holds for every output norm. Under a score budget, squaring the Euclidean norm gives an exponential polynomial with a controlled positive envelope. At the same score budget, a different fixed norm can increase the number of threshold crossings.

Define one norm on $\mathbb{R}^2$ by

$$c_0 = \log\log(e+1), \qquad \mathcal{A} = \{1\} \cup \left\{ 1 + \frac{c_0}{\log\log(n+e)} : n \geq 1 \right\},$$
$$\|(u,v)\|_* = \sup_{a \in \mathcal{A}} \left| \frac{2u}{a} - \frac{v}{a^2} \right|. \tag{20}$$

The supremum is finite, and the functionals for $a = 1, 2$ are linearly independent, so this is a norm. Neither the dimension nor the norm depends on the gain precision $\delta$. Write $\mathcal{C}_*^{B,L}(H,\delta)$ for the capacity defined in Section 5 with this fixed norm in place of the Euclidean norm.

**Proposition 7** (Fixed-norm capacity separation)**.** *For the norm in (20), two finite attention heads with scalar payload, $B = 2$, and $L = \log 2$ guarantee $2N$ transitions for every integer $N \geq 1$ satisfying $M_N(\delta) < 1 - 3\delta$, with $a_n = 1 + c_0/\log\log(n+e)$ and*

$$M_N(\delta) = \max_{n \leq N} \sup_{a \in \mathcal{A}} \left[ \frac{2q_n}{a} - \frac{q_n^2}{a^2} + \delta \left( \frac{2q_n}{a} + \frac{q_n^2}{a^2} \right) \right], \qquad q_n = (a_n + a_{n+1})/2.$$

*Consequently, their robust capacity satisfies*

$$\mathcal{C}_*^{2,\log 2}(2,\delta) = \Omega\left( \frac{\delta^{-1/2}}{D(\log D)^2} \right), \qquad D = \log(1/\delta), \quad \text{as } \delta \downarrow 0. \tag{21}$$

*The Euclidean capacity at the same head count and score budget is $\Theta(D/\log D)$.*

The construction follows the parabola $y = (q, q^2)$ for $q = e^z \in [1, 2]$. Its supporting functional at $a \in \mathcal{A}$ satisfies

$$\ell_a(q, q^2) = \frac{2q}{a} - \frac{q^2}{a^2} = 1 - (q/a - 1)^2.$$

At $a_n = 1 + c_0/\log\log(n+e)$ the norm touches one; between successive contacts its deficit at $q_n$ is exactly $(a_n - a_{n+1})^2/(4a_n^2)$. Since $a_n - a_{n+1} \asymp 1/[n \log n (\log\log n)^2]$, these contacts yield (21). Appendix J gives the three-token finite-attention realization and gain-box margins. Equivalence of norms does not preserve a crossing count with shrinking relative margins.

## 8 Related Work

**Head count, rank, and approximation complexity.** Bhojanapalli et al. (2020) identify a low-rank bottleneck caused by coupling the number of heads to the dimension of each head. Amsel et al. (2025) give a nearest-neighbor task representable by one full-rank head, while low-rank heads require a head count exponential in the embedding dimension. Yu et al. (2026) study a generalized retrieval task and derive head-dependent parameter requirements for approximation. These results show why head count must be interpreted together with the resources allocated to each head. Our capacity bounds isolate head-count dependence under independent gain robustness, with arbitrary finite width and context size. This dependence persists even after two heads approximate every smooth scalar response in the chosen interface.

**Discrete computation and head hierarchies.** Hahn (2020) establish limitations of fixed-size attention architectures for formal-language tasks as sequence length grows. More directly, Rajaraman et al. (2026) prove a strict parity hierarchy in a single attention-only layer: three heads compute three-bit parity, whereas two cannot, even with unbounded embedding dimension and numerical precision. Their result establishes a head-count advantage for exact computation of a Boolean task family. Their denominator-clearing proof uses an alternating-sum obstruction on the Boolean cube. We instead bound maximal robust oscillation over finite attention realizations of a localized source effect along a continuous affine query path. Two-head smooth approximation is already dense in this interface; every counted threshold label must nevertheless survive independent head gains. Clearing positive softmax denominators now gives an exponential polynomial whose gain-box margins and spectral-envelope curvature yield bounds in $H$, the relative-score budget $BL$, and gain precision $\delta$.

**Universality and positive exponential representations.** Yun et al. (2020) establish Transformer universality for continuous sequence-to-sequence functions using attention and feed-forward layers. Kajitsuka & Sato (2024) show that a single low-rank self-attention layer can encode sequence context and, together with two feed-forward networks, gives universal approximation on compact domains. These architectures use nonlinear feed-forward computation in addition to attention. Our density theorem establishes smooth density for a scalar intervention response with a linear observed readout: two reciprocal softmax denominators give $C^r$ approximation along a fixed affine query chart. Related positive exponential constructions establish density for mixture-of-experts mean functions (Nguyen et al., 2016) and differences of log-sum-exp networks (Calafiore et al., 2020). Here coefficientwise positivity also constrains the robust lower constructions: rational amplification and the square-envelope correction must preserve both denominator positivity and the full independent-gain margin.

**Memorization and statistical capacity.** Mahdavi et al. (2024) prove that an $H$-head attention layer can memorize $\Omega(Hn)$ examples under specified input-independence conditions, where $n$ is context length. Their construction uses softmax saturation to allocate examples across heads. Edelman et al. (2022) show that bounded-norm attention can represent sparse input dependencies with sample complexity logarithmic in context length. Robust alternating capacity fixes an ordered query path and counts labels certified over every admissible head gain. It therefore measures a different resource from storing arbitrary labeled sequences or bounding statistical sample complexity. In particular, our unrestricted upper bound is independent of context length and head-map magnitudes, and the score-dependent bounds retain this independence while resolving the joint $H, BL, \delta$ scaling.

**Head pruning and perturbation stability.** Michel et al. (2019) find that many trained attention heads can be removed with little change in task performance, while Voita et al. (2019) identify specialized heads and study which survive pruning. Head removal sets a gain to zero; our box $[1-\delta, 1+\delta]^H$ instead tests all simultaneous relative rescalings around a fixed decomposition. This gives a continuous stability measure for the contributions exposed by that decomposition. Tsai et al. (2021) analyze classification margins and generalization under neural-network weight perturbations. Our gain perturbations leave keys, queries, and scores fixed, allowing a sharp capacity calculation that isolates cancellation between heads.

**Smoothness and oscillation counts.** Kim et al. (2021) show that dot-product attention is not globally Lipschitz on unbounded input domains and analyze a Lipschitz alternative. Castin et al. (2024) quantify the effects of sequence length, bounded inputs, and masking on attention's Lipschitz constant. We bound relative score slopes and count robust threshold transitions, so the target quantity is a crossing count rather than an input–output Lipschitz constant. Crossing numbers also drive neural-network depth separations: Telgarsky (2016) compares oscillations of deep and shallow networks along affine lines. Our count requires each selected label to survive an entire head-gain box, while width and the number of exponential terms remain unrestricted. The truncated potential controls these robust crossings for arbitrary output norms; the tilted complex-envelope argument supplies the sharper Euclidean score law. The fixed-norm separation identifies precisely why threshold geometry enters this latter step.

## 9 Conclusion

Smooth density saturates at two heads, while robust decision capacity scales with head count. For the localized source interface, every finite $H$-head layer obeys an unrestricted robust-alternation ceiling of order $H\delta^{-1}\log(1/\delta)$, and finite layers attain that order. Under a score budget, matching bounds give a factorial regime followed by the $\sqrt{HBL\log(1/\delta)}$ regime; fixed-head high-score laws connect the bounded and unrestricted scales. These results characterize the head–score resource tradeoff for gain-stable decisions in the localized source interface. The unrestricted upper bound holds for any output norm. The score-dependent rates apply to Euclidean output geometry; the fixed-norm separation shows that changing the output norm can change their order.

### AI USE STATEMENT

Generative AI tools were used during manuscript preparation to assist with drafting and language editing, including improving the clarity, grammar, and phrasing of the text. They were also used to support literature search and identify potentially relevant prior work. We manually verified the related work suggested by AI tools and ensured the accuracy of the corresponding citations. We retain full responsibility for the manuscript, including its scientific claims, mathematical arguments, and cited references.

### REPRODUCIBILITY STATEMENT

The mathematical setting, intervention interface, gain perturbations, and capacity definitions are specified in Section 2. Appendix A gives finite attention realizations and the smooth density proof. Appendix B proves the unrestricted capacity bounds, and Appendices C and D give the score-constrained upper bounds and finite lower constructions. The high-score construction and its context budget appear in Appendix E, with the matching upper bound in Appendix F. Appendix J provides the explicit output-norm separation.

## REFERENCES

Noah Amsel, Gilad Yehudai, and Joan Bruna. Quality over quantity in attention layers: When adding more heads hurts. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9c537882044c8b5352c363e840872ddb-Abstract-Conference.html.

Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 864–873. PMLR, 2020. URL https://proceedings.mlr.press/v119/bhojanapalli20a.html.

Giuseppe C. Calafiore, Stephane Gaubert, and Corrado Possieri. A universal approximation result for difference of log-sum-exp neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 31(12):5603–5612, 2020. doi: 10.1109/TNNLS.2020.2975051. URL https://doi.org/10.1109/TNNLS.2020.2975051.

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5817–5840. PMLR, 2024. URL https://proceedings.mlr.press/v235/castin24a.html.

Benjamin L. Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5793–5831. PMLR, 2022. URL https://proceedings.mlr.press/v162/edelman22a.html.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020. doi: 10.1162/tacl_a_00306. URL https://aclanthology.org/2020.tacl-1.11/.

Tokio Kajitsuka and Issei Sato. Are transformers with one layer self-attention using low-rank weight matrices universal approximators? In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/96ee35c170cb8720033b16259c305da9-Paper-Conference.pdf.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz constant of self-attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5562–5571. PMLR, 2021. URL https://proceedings.mlr.press/v139/kim21i.html.

Sadegh Mahdavi, Renjie Liao, and Christos Thrampoulidis. Memorization capacity of multi-head attention in transformers. In *International Conference on Learning Representations*, 2024. URL

https://proceedings.iclr.cc/paper_files/paper/2024/hash/d8542126cd3e0dd6c0a44e0aa1957072-Abstract-Conference.html.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/2c601ad9d2ff9bc8b282670cdd54f69f-Paper.pdf.

Hien D. Nguyen, Luke R. Lloyd-Jones, and Geoffrey J. McLachlan. A universal approximation theorem for mixture-of-experts models. *Neural Computation*, 28(12):2585–2593, 2016. doi: 10.1162/NECO_a_00892. URL https://doi.org/10.1162/NECO_a_00892.

Rajmohan Rajaraman, Ravi Sundaram, and Amanuel Tesfaye. The head complexity of Boolean functions in single-layer attention. arXiv preprint arXiv:2609.04046, 2026. URL https://arxiv.org/abs/2609.04046.

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

Yu-Lin Tsai, Chia-Yi Hsu, Chia-Mu Yu, and Pin-Yu Chen. Formalizing generalization and adversarial robustness of neural networks to weight perturbations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 19692–19704. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a3ab4ff8fa4deed2e3bae3a5077675f0-Paper.pdf.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580/.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001. URL https://doi.org/10.1561/2200000001.

Penghao Yu, Haotian Jiang, Zeyu Bao, Ruoxi Yu, and Qianxiao Li. The effect of attention head count on transformer approximation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7d72a514fc6948515af3ee69c1059776-Abstract-Conference.html.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=ByxRM0Ntvr.

## A Finite Attention Realization and Smooth Density

**Implementing reciprocal positive exponential sums.** For each head, consider $2 + \sum_j e^{a_{hj}z+b_{hj}}$. Give the source token and final query token score zero. Use independent coordinates for token identities, query control, and source payload. Take query input $x_q = e_q + ze_{\mathrm{ctrl}}$, the source is $x_s = e_s + ve_{\mathrm{pay}}$, and background tokens are $x_j = e_j$. Choose a head dimension $d_h \geq 2$, and set $W_h^Q e_{\mathrm{ctrl}} = f_1, \quad W_h^Q e_q = f_2, \quad W_h^K e_j = \sqrt{d_h}(a_{hj}f_1 + b_{hj}f_2)$. All other key actions vanish, including on source, query, control, and payload coordinates. The final-query scaled dot products therefore give the prescribed affine scores, invariant under payload intervention. Set $W_h^V e_{\mathrm{pay}} = f_1$, all other value actions to zero, and project this coordinate to the desired scalar or vector coefficient. Padding model and head dimensions accommodates any finite token set; the query residual cancels from the intervention difference.

For a positive finite exponential sum $Q_h$ with a positive constant atom, choose a common $\zeta > 0$ below half every constant coefficient. Split off source and query atoms of weight $\zeta$; the rest are background tokens. Dividing weights by $\zeta$ sets source and query scores to zero and gives source weight $\zeta/Q_h$. Readout $b_h/\zeta$ realizes $b_h/Q_h$ with unchanged source-relative slopes. Use the union of all head-specific atoms as a common context, assigning sufficiently small positive constant weights to unused atoms. Strict witness margins in the constructions below preserve every label and slope bound under these additions. Zero-readout heads pad the construction to $H$ heads.

**Completing the density proof.** On the positive interval $e^I$, the function $f(\log y)$ is $C^r$. Uniformly approximate its $r$th derivative by a polynomial, then integrate $r$ times with lower derivatives matched at one endpoint. Composing this $C^r$ approximation with $e^z$ gives $R(z) = \sum_{k=0}^n r_k e^{kz}$ in the same topology on $I$. For arbitrary $\xi_k > 0$, take $A_\pm(z) = \sum_{k=0}^n (\xi_k + \max\{\pm r_k, 0\})e^{kz}$. Then $A_+ - A_- = R$. As positive $\lambda$ tends to zero, the denominators in (5) stay away from zero and converge to 2 in $C^r$. Differentiating their reciprocals gives $g_\lambda \to R$ in $C^r(I)$. The finite realization with fixed query chart $x_q = e_q + ze_{\mathrm{ctrl}}$ proves Theorem 2.

## B Proof of the Unrestricted Capacity Law

**Upper bound.** Write $c_h = \|B_h\|_{\mathrm{op}}$ and $a_h(z) = c_h w_h(z)$. At an inactive witness, choose a unit payload $v$ and a dual-unit output functional $\psi$ with $\psi(B_h v) = \|B_h\|_{\mathrm{op}}$, and put $\ell(M) = \psi(Mv)$. For $A = a_h(z)$ and signed sum $s$ of the other projected contributions, the gain-box support function with $0 < \delta \leq 1$ gives

$$\varepsilon \geq |A + s| + \delta \left( A + \sum_{j \neq h} |\ell(B_j)| w_j \right) \geq \delta(|A + s| + A + |s|) \geq 2\delta A.$$

Consequently

$$a_h(z) \leq C := \varepsilon/(2\delta). \tag{22}$$

Unimodality of $a_h$ makes $\{z : a_h(z) > C\}$ an interval. It contains no inactive witness and at most one active witness in an alternating sequence. Deleting these witnesses across heads costs at most $2H$ adjacent edges; all retained endpoints obey (22) for every head.

For active and inactive endpoints $p$ and $q$ of a retained edge, the compact convex gain-response set at $p$ avoids the closed $\varepsilon$-ball in the finite-dimensional space of payload-to-output maps, equipped with the operator norm. Strict separation in this space gives a dual-unit functional $\ell$, which need not have the preceding form $\psi(Mv)$, with

$$\inf_t \ell\left( \sum_h t_h w_h(p) B_h \right) > \varepsilon, \qquad \sup_t \ell\left( \sum_h t_h w_h(q) B_h \right) \leq \varepsilon.$$

Put $\alpha_h = \ell(B_h)$ and $b_h = |\alpha_h| \leq c_h$. Expanding the gain-box extrema yields

$$\sum_h b_h |w_h(p) - w_h(q)| > \delta \sum_h b_h (w_h(p) + w_h(q)).$$

Put $S = \sum_h b_h(w_h(p) + w_h(q))$. Active separation implies $\sum_h b_h w_h(p) > \varepsilon/(1-\delta)$, so $S > \varepsilon$. For any $\tau > 0$, let $\Phi(a) = \log(1 + a/\tau)$. For $M \geq m \geq 0$, $\Phi(cM) - \Phi(cm)$ is nondecreasing in $c \geq 0$. Using $\log(x/y) \geq 2(x-y)/(x+y)$ for $x \geq y > 0$, and $c_h \geq b_h$, gives

$$\sum_h |\Phi(a_h(p)) - \Phi(a_h(q))| \geq \frac{2\sum_h b_h |w_h(p) - w_h(q)|}{2\tau + S} > \frac{2\delta S}{2\tau + S} > \frac{2\delta}{1 + 2\tau/\varepsilon}. \tag{23}$$

Although the functional depends on the edge, (23) does not.

For any $\tau > 0$, take $y_h(z) = \Phi(\min\{a_h(z), C\})$. This function is unimodal, takes values in $[0, \log(1 + C/\tau)]$, and agrees with $\Phi(a_h)$ at every retained endpoint. Hence $\mathrm{TV}_I(y_h) \leq 2\log(1 + C/\tau)$ and

$$K \leq 2H + \frac{H}{\delta}(1+r)\log\left(1 + \frac{1}{\delta r}\right), \qquad r = 2\tau/\varepsilon > 0.$$

Taking the infimum proves Theorem 3's upper bound. For its asymptotic expansion, set $u = 1/r$. The elementary lower bound $(1+1/u)\log(1+u/\delta) \geq D + \log u + D/u$ for $u \geq 1$ is minimized at $u = D$; the same choice gives the matching upper expansion $D + \log D + O(1)$. For $u < 1$, write $v = u/\delta$; if $v \leq \delta^{-1/2}$, the factor $\log(1+v)/u$ exceeds this scale, and otherwise $(1 + 1/u)\log(1 + u/\delta) \geq D + \log u + D/(2u)$, again giving $D + \log D - O(1)$. For $H = 1$, nonnegativity and unimodality forbid active–inactive–active reentry and more than two transitions.

**Matching finite construction.** Fix $0 < \delta \leq 1/8$ and threshold $\varepsilon > 0$. Put $Q_\delta = \frac{\log((1-\delta)/\delta)}{\log((1+\delta)/(1-\delta))}$, $n = \lceil Q_\delta \rceil - 1$. Since $n < Q_\delta$, continuity permits $a \in (\delta, 1)$ with $n < \log(2(1-a)/(a+\delta))/\log((1+a)/(1-a))$. Set $r = (1-a)/(1+a)$ and $T_i = 2\varepsilon r^i/(a+\delta)$. Then $T_i > \varepsilon/(1-a)$ for $0 \leq i \leq n$, including when $Q_\delta$ is an integer. Define $g_i = \varepsilon + (-1)^i a T_i$, $u_i = \frac{T_i + g_i}{2}$, $v_i = \frac{T_i - g_i}{2}$. Both $u_i, v_i$ are positive, and $T_i - T_{i+1} = a(T_i + T_{i+1}) = |g_i - g_{i+1}|$ makes them nonincreasing. Under gains $t_+, t_-$, the target response satisfies $t_+ u_i - t_- v_i \in [g_i - \delta T_i, g_i + \delta T_i]$. For even $i$, the minimum $\varepsilon + (a-\delta)T_i > \varepsilon$ gives strict activity. For odd $i$, the maximum $\varepsilon - (a-\delta)T_i < \varepsilon$ and minimum $\varepsilon - (a+\delta)T_i > -\varepsilon$ (since $T_i < T_0$) give strict inactivity.

Choose output coefficients $c_+ = 2u_0$ and $c_- = 2v_0$ and target denominators $D_\pm(i) = c_\pm/(u_i \text{ or } v_i)$. Their increments $d_{\pm,j} = D_\pm(j) - D_\pm(j-1)$ are nonnegative. For finite $R > 1$, define

$$P_{\pm,R}(z) = 2 + \sum_{j=1}^{n} d_{\pm,j}\left[e^{R^j(z-1)} + e^{R^j(-z-1)}\right], \qquad z_i = 1 - R^{-i-1/2}.$$

At the finite witness set, $P_{\pm,R}(\pm z_i) \to D_\pm(i)$ as $R \to \infty$: the first exponential at $z_i$ tends to one for $j \leq i$ and zero for $j > i$, while the reflected term vanishes. For sufficiently large finite $R$, the response $\frac{c_+}{P_{+,R}(z)} - \frac{c_-}{P_{-,R}(z)}$ preserves all robust labels. The ordered witnesses $-z_n, \ldots, -z_0, z_1, \ldots, z_n$ give $2n$ transitions with the same two heads and at most $2n$ background atoms. Omit zero increments and assign unused atoms small positive weights as in Appendix A; all logits are finite.

For $J = \lfloor H/2 \rfloor$, place $J$ copies in disjoint query subintervals. Two zero-value sentinel tokens per pair, with logits $M(a_r - z)$ and $M(z - b_r)$, suppress that pair outside its assigned interval. Foreign tokens receive constant logit $-L$. Finiteness of the witness set lets finite $M, L$ preserve all strict margins. Concatenation gives at least $2Jn = 2\lfloor H/2 \rfloor(\lceil Q_\delta \rceil - 1)$ transitions, exactly the lower guarantee in (6). Since $2Q_\delta = \delta^{-1}D - 1 + O(\delta D)$, this guarantee is $\lfloor H/2 \rfloor[\delta^{-1}D + O(1)]$, with a uniform error per head pair.

**Optimal coefficient for two scalar heads.** For $0 < \delta \leq 1/8$, put $\rho_\delta = (1+\delta)/(1-\delta)$ and use $Q_\delta$ above. We prove the finite bounds

$$2\lceil Q_\delta \rceil - 2 \leq \mathcal{C}^{\mathrm{sc}}_{\mathrm{attn}}(2, \delta) \leq 2\lceil Q_\delta \rceil + 6. \tag{24}$$

The preceding construction with $H = 2$ gives the lower bound.

For the upper bound with opposite coefficient signs, write the response as $t_+ u - t_- v$, where $u, v \geq 0$ are unimodal, and set $T = u + v$, $g = u - v$. The active and inactive conditions are respectively $|g| > \varepsilon + \delta T$ and $|g| \leq \varepsilon - \delta T$. Split the witness interval at the two modes. On the outer branches,

$u$ and $v$ change in the same direction, so adjacent active/inactive $p, q$ satisfy $|T(p) - T(q)| \geq |g(p) - g(q)| \geq \big||g(p)| - |g(q)|\big| > \delta(T(p) + T(q))$. Consequently $|\log T(p) - \log T(q)| > \log \rho_\delta$. Every active witness has $T > \varepsilon/(1-\delta)$, and every inactive witness has $T \leq \varepsilon/\delta$. Monotonicity of $T$ permits at most one witness below and one above this range, both at branch endpoints. Deleting them costs at most two edges; the remaining $k$ obey $k < Q_\delta$, giving at most $\lceil Q_\delta \rceil + 1$ per outer branch. On the middle branch, one amplitude increases while the other decreases. The two inactive inequalities $(1+\delta)u - (1-\delta)v \leq \varepsilon$ and $(1+\delta)v - (1-\delta)u \leq \varepsilon$ have opposite monotonicities, so their intersection is an interval and contributes at most two alternating edges. At most two further edges cross branch boundaries. This proves the upper bound in (24) for opposite signs.

For equal signs, the response magnitude is $t_+u + t_-v$. At an inactive witness $u+v \leq C_0 := \varepsilon/(1+\delta)$. Deleting the active witnesses where either amplitude exceeds $C_0$ removes at most four edges. On retained edges the change in $u+v$ exceeds $2\delta\varepsilon/(1-\delta^2)$, whereas the combined total variation of $\min(u, C_0)$ and $\min(v, C_0)$ is at most $4C_0$. Hence $K < 4 + 2(1-\delta)/\delta$, which is smaller than $2\lceil Q_\delta \rceil + 6$: for $\delta \leq 1/8$, $\log((1-\delta)/\delta) \geq 2/(1+\delta)$ and $\log \rho_\delta \leq 2\delta/(1-\delta^2)$ imply $Q_\delta \geq (1-\delta)/\delta$. Zero coefficients are included in this case. Finally, $2Q_\delta = \delta^{-1}\log(1/\delta) - 1 + O(\delta \log(1/\delta))$, so (24) proves Corollary 4.

## C Upper Bounds under a Score Budget

**Lemma S1** (Spectral oscillation bound)**.** *Let $F$ be a nonzero real exponential polynomial and $A$ a positive exponential sum dominating its absolute coefficients, so $|F(z)| \leq A(\operatorname{Re} z)$. Suppose $A$ has exponent diameter at most $W$, and $F$ alternates signs at witnesses over length $L$. If each negative witness $p$ satisfies $|F(p)| \geq \rho A(p)$, then*

$$K \leq \inf_{N \in \mathbb{N}} \left\{ N - 1 + \frac{N \log(1/\rho)}{\log 2} + \frac{12WL}{N \log 2} \right\}. \tag{25}$$

*In particular, $K = O(d + \sqrt{WLd})$ for $d = 1 + \log(1/\rho)$, uniformly for $0 < \rho \leq 1$.*

*Proof.* Write $A(z) = \sum_j a_j e^{\lambda_j z}$, $\phi = \log A$, and $\pi_z(j) = a_j e^{\lambda_j z}/A(z)$. Then $\partial_z \log \pi_z(j) = \lambda_j - \phi'(z)$, giving Fisher information $\mathcal{I}_A = \phi'' = \operatorname{Var}_{\pi_z}(\lambda)$ and $\int_{\mathbb{R}} \mathcal{I}_A = \lambda_{\max} - \lambda_{\min} \leq W$. The KL divergence satisfies $D_{\mathrm{KL}}(\pi_p \| \pi_t) = \phi(t) - \phi(p) - \phi'(p)(t-p)$. Partition the witnesses into $N$ equal cells of length $\ell = L/N$, assigning shared endpoints once. In a cell with a sign transition, choose a negative witness $p$ and tilt $f(z) = e^{-\phi'(p)(z-p)} F(z)$. Taylor's integral remainder gives

$$\log \frac{\max_{|z-p|=2\ell} |f(z)|}{|f(p)|} \leq \log(1/\rho) + 2\ell \int_{p-2\ell}^{p+2\ell} \phi''(u)\, du.$$

The zero count in $|z-p| \leq \ell$, with multiplicity, is at most this bound divided by $\log 2$. Indeed, divide by the Blaschke factors of these zeros in the radius-$2\ell$ disk. Each has boundary modulus one and center modulus at most $1/2$, so the maximum principle gives $|f(p)| \leq M 2^{-n}$ for $n$ zeros and boundary maximum $M$. Internal transitions give distinct zeros. The expanded intervals overlap at most six times, and connecting cells costs at most $N-1$ transitions. Summing proves (25) for every $N$. Writing $a = 1 + \log(1/\rho)/\log 2$ and $b = 12WL/\log 2$, the minimizing integer is among $\max\{1, \lfloor\sqrt{b/a}\rfloor\}$ and $\max\{1, \lceil\sqrt{b/a}\rceil\}$, since the objective is $aN + b/N - 1$. This also gives the stated asymptotic bound. □

**Application to Euclidean attention.** With the cleared $N(z), Q(z)$ of Section 5, put

$$F = \|N\|_2^2 - \varepsilon^2 Q^2, \qquad A = \varepsilon^2 Q^2 + \left( \sum_h \|b_h\|_2 Q/Q_h \right)^2.$$

Here $\|N\|_2^2$ is the holomorphic continuation of its real exponential-polynomial expansion. The positive $A$ dominates $F$'s absolute coefficients and has exponent diameter at most $4HB$. At an inactive witness, independent uniform signs $\sigma_h$ give $\mathbb{E}\|\sum_h (1+\delta\sigma_h) b_h/Q_h\|_2^2 \leq \varepsilon^2$. Expansion

and Cauchy–Schwarz bound $A \le \varepsilon^2 Q^2(1 + H/\delta^2)$. The common gain $1 + \delta$ gives the exact nominal deficit $-F \ge [1 - (1 + \delta)^{-2}]\varepsilon^2 Q^2$. Thus (25) applies with

$$\rho = \frac{1 - (1 + \delta)^{-2}}{1 + H/\delta^2} = \frac{\delta^3(2 + \delta)}{(1 + \delta)^2(H + \delta^2)}.$$

For $H \ge 2$ and $0 < \delta \le 1/8$, $d = 1 + \log(1/\rho) = \Theta(D + \log(H + 1)) = \Theta(U)$. The lemma gives (9).

**The factorial refinement.** Discard an active first witness, leaving $m \ge K - 1$ transitions from inactive $z_0$. Multiply $F, A, Q^2$ by $e^{2HBz}$ without renaming: the exponents lie in $[0, W]$, $W = 4HB$, and $Q^2$ is nondecreasing. Common gains $1 \pm \delta$ imply $|F(z_i)| \ge [1 - (1 + \delta)^{-2}]\varepsilon^2 Q^2(z_i) \ge \rho A(z_0)$ at every witness; the envelope bound is needed only at $z_0$. The order-$m$ divided-difference summands have one sign and denominator magnitudes at most $L^m$. The mean-value formula gives

$$\frac{(m + 1)\rho A(z_0)}{L^m} \le \frac{\sup_I |F^{(m)}|}{m!} \le \frac{W^m e^{WL} A(z_0)}{m!}.$$

Hence, for $X = WL = 4Hx$, the exact finite constraint is

$$\log \Gamma(m + 2) - m \log X \le X + \log(1/\rho). \tag{26}$$

The feasible set is finite because $(m + 1)!/X^m \to \infty$. Thus, before inversion,

$$K \le 1 + \max\{m \in \mathbb{Z}_{\ge 0} : \rho(m + 1)! \le e^X X^m\}.$$

Stirling inversion yields $K = O(1 + X + d/\log(e + d/X))$, $d = 1 + \log(1/\rho)$. For $H \le D$, $d = \Theta(D)$. When $Hx \le D$, the $X$ term is absorbed because $t \log(e + 1/t) = O(1)$ for $0 < t \le 1$. This gives the first upper branch of (10); the spectral estimate gives the second.

## D Matching Finite Constructions under a Score Budget

**An exact rational-amplification certificate.** Let $A$ be a finite positive exponential sum with a positive constant atom, and let $P$ have coefficientwise $|P_j| \le A_j$. Suppose $r = P/A$ makes $m$ traversals of $[-a, a]$, $0 < a \le 1$, and put $R = A_{\max}/A_{\min}$. For $k \ge 2$, set $\theta_j = (2j - 1)\pi/(2k)$, $c_j = (1 + \cos\theta_j)/4$, and $Q_j = A + c_j P$. These denominators have positive coefficients on $A$'s exponent support. Partial fractions give

$$\frac{T_{k-1}(r/a)}{\prod_j (1 + c_j r)} = \sum_j \frac{\beta_j}{1 + c_j r}, \qquad \beta_j = \frac{c_j^{k-1} T_{k-1}(-1/(ac_j))}{\prod_{l \ne j}(c_j - c_l)}.$$

Define the exact quantities $P_k(a) = \prod_j (1 + c_j a)$ and $S_k(a) = \sum_j |\beta_j|/(1 - c_j)$. At $r = a\cos(i\pi/(2(k - 1)))$, the response $g = \sum_j \beta_j/Q_j$ has magnitude at least $1/(A_{\max} P_k(a))$ for even $i$, and zero for odd $i$. Its gain uncertainty throughout the path is at most $\delta S_k(a)/A_{\min}$. With threshold $\varepsilon = 1/(2A_{\max} P_k(a))$, the exact sufficient condition

$$2\delta R P_k(a) S_k(a) \le \gamma, \qquad 0 < \gamma < 1, \tag{27}$$

gives normalized label slack at least $1 - \gamma$ and head amplitudes at most $\gamma/\delta$. Ordered level-hitting points exist by continuity; removing repeated active boundary labels gives $2m(k - 1)$ transitions. Appendix A implements the denominators as finite attention.

For the asymptotic budget, the node polynomial $2^{1-3k} T_k(4c - 1)$ has derivative magnitude $k/(8^{k-1} \sin\theta_j)$ at $c_j$. Therefore

$$|\beta_j| = \frac{(8c_j)^{k-1} \sin\theta_j}{k} \cosh\left((k - 1)\,\mathrm{arcosh}\,\frac{1}{ac_j}\right).$$

Since $\mathrm{arcosh}\, u \le \log(2u)$, this implies $P_k(a) \le (1 + a/2)^k$ and $S_k(a) \le 2(16/a)^{k-1}$. Thus a fully explicit, conservative logarithmic budget is $\log R + \log(4/\gamma) + k \log(1 + a/2) + (k - 1)\log(16/a)$; the exact certificate uses the product and residue sum themselves.

**The factorial phase.** Let $1 \le x$, $Hx \le D$, and use the full budget $B = x/L$. For an integer $m \ge 1$, set $y = e^{Bz/m}$, $t(y) = [2y - (1 + e^{x/m})]/(e^{x/m} - 1)$, $P = T_m(t(y)) = \sum_{j=0}^{m} p_j y^j$, and $A = \sum_{j=0}^{m} |p_j| y^j$. The positive roots of $P$ make its coefficients nonzero with alternating signs. Thus $A$ has a positive constant atom and slopes in $[0, B]$. Writing $u_x = 3 + 4/(e^{x/m} - 1)$ gives the exact maximum $E = A(L) = \cosh(m \operatorname{arcosh} u_x)$, while $R \le e^x$. At successive extrema $P = \pm 1$, $P/A$ makes $m$ traversals of $[-1/E, 1/E]$. Hence (27), with $a = 1/E$, is a finite certificate before any asymptotic simplification. The exact log envelope is $\log\cosh(m \operatorname{arcosh} u_x)$, which is at most $m \operatorname{arcosh} u_x$. For every $x > 0$, define the finite feasible set

$$\mathcal{F}_C = \{(k, m) \in \mathbb{N}^2 : 2 \le k \le H,\ m \ge 1, \quad 2\delta e^x P_k(1/E) S_k(1/E) < 1\}.$$

Every member guarantees $2m(k-1)$ transitions, so $K_C = \max(\{0\} \cup \{2m(k-1) : (k, m) \in \mathcal{F}_C\})$ is an integer lower bound. The maximum exists by the finite spectral upper bound. This certificate also applies to the vanishing-score range.

Put $A_0 = \log(e + D/(Hx))$ and $q = D/A_0$. Fix $\gamma \in (0, 1)$. For a sufficiently small absolute $c > 0$, choose $M = \lfloor cq \rfloor$, $k = \min\{H, M + 1\}$, and $m = \lfloor M/(k-1) \rfloor$ whenever $M \ge 1$. Then $m(k-1) = (k-1)\lfloor M/(k-1) \rfloor \in [M/2, M]$ and $m \le 1 + 2cD/H$. For $D/(Hx) \ge 1$, $\log(e + m/x) = O(A_0)$, so the budget is $x + O(MA_0 + 1) \le D/2 + O(cD + 1)$. Choosing $c$ small certifies $\Omega(q)$ transitions for all sufficiently large $D$. Bounded $q$ is covered by one transition. The remaining bounded-$D$ region $D \ge 32$, $2 \le H \le D$, $1 \le x \le HD$ has bounded target rate and also admits one transition, since $x \ge 1 > \log((1 + \delta)/(1 - \delta))$. Thus the lower constant is uniform. In particular, $x = D/H$ gives $\Omega(D)$ transitions with the same amplitude and slack guarantees.

**An exact finite Gaussian lattice.** For $0 < \eta \le 1$ and even $m \ge 2$, put $p = \lceil \pi/\eta \rceil$ and $s = mz/L$. Define

$$A(s) = \sum_{j=-p}^{m+p} e^{\eta(2js - j^2 - ms)}, \qquad P(s) = \sum_{j=-p}^{m+p} (-1)^j e^{\eta(2js - j^2 - ms)}.$$

The atom $j = m/2$ is constant, and the exact score cost is

$$x_0 = \eta m(m + 2p) = \eta m^2 + 2\pi m + r_{\eta,m}, \qquad r_{\eta,m} = 2\eta m(\lceil \pi/\eta \rceil - \pi/\eta) \in [0, 2\eta m). \tag{28}$$

Write $\Theta_\eta = \sum_{j \in \mathbb{Z}} e^{-\eta j^2}$, $C_\eta = \sum_{j \in \mathbb{Z}} (-1)^j e^{-\eta j^2}$, and $E_{\eta,p} = 2\sum_{j=p+1}^{\infty} e^{-\eta j^2}$. The periodized Gaussian is maximal at an integer; the nearest lattice point is within $1/2$. Consequently

$$A_{\max} \le \Theta_\eta, \qquad A_{\min} \ge e^{-\eta(m^2+1)/4}, \qquad R \le \Theta_\eta e^{\eta(m^2+1)/4}. \tag{29}$$

The Gaussian Fourier transform, obtained by integration by parts and its value at zero, gives

$$C_\eta = 2\sqrt{\pi/\eta} \sum_{l \ge 0} e^{-\pi^2(2l+1)^2/(4\eta)}.$$

The integral tail bound $E_{\eta,p} \le e^{-\eta p^2}/(\eta p)$ shows $E_{\eta,p} < C_\eta$ for $0 < \eta \le 1$. At integer core witnesses, therefore, $(-1)^i P(i)/A(i) \ge a_{\eta,p} := (C_\eta - E_{\eta,p})/\Theta_\eta > 0$. There are $m$ traversals of $[-a_{\eta,p}, a_{\eta,p}]$. Uniformly for $0 < \eta \le 1$, $\Theta_\eta = \Theta(\eta^{-1/2})$ and $\log(1/a_{\eta,p}) = \pi^2/(4\eta) + O(1)$.

Combining these expressions with the rational certificate gives $K \ge 2m(k-1)$ whenever

$$\begin{gathered} 2 \le k \le H, \quad m \ge 2 \text{ even}, \quad 0 < \eta \le 1, \quad \eta m(m + 2\lceil \pi/\eta \rceil) \le x, \\ \log(2/\gamma) + \log \Theta_\eta + \eta(m^2 + 1)/4 + \log P_k(a_{\eta,p}) + \log S_k(a_{\eta,p}) \le D. \end{gathered} \tag{30}$$

For fixed $\eta$, the largest score-feasible even integer is $m_{\text{score}} = 2\lfloor (\sqrt{p^2 + x/\eta} - p)/2 \rfloor$. Check the gain condition for each even $2 \le m \le m_{\text{score}}$ and $2 \le k \le H$. Maximizing $2m(k-1)$ over these choices and $0 < \eta \le 1$ gives an integer lower guarantee $K_G$ (zero if none are feasible), with no unspecified parameter-selection constants. The sum has exactly $m + 2p + 1$ distinct exponents. Splitting its constant coefficient between source and query gives a common context of exactly $m + 2p + 2$ tokens.

**The square-root phase up to $x = D$.** Suppose $2 \le H \le D$, $1 \le x \le D$, and $Hx \ge D$. Fix $\gamma \in (0,1)$. Choose a sufficiently large absolute $C_0$ and then a sufficiently small $c_0 > 0$ with $C_0 c_0 < 1$. For sufficiently large $D$, take $k = \min\{H, \max\{2, \lfloor c_0 H \rfloor\}\}$, $\eta = C_0 k/D$, $p = \lceil \pi/\eta \rceil$, and the largest even integer $m$ satisfying $\eta m(m+2p) \le x$. The certificate's logarithmic budget is

$$\frac{\eta m^2}{4} + \frac{\pi^2(k-1)}{4\eta} + O(k + \log(1/\eta)) \le \frac{D}{4} + \frac{\pi^2 D}{4C_0} + O(c_0 D + \log D) < D.$$

The constants are chosen in this order to make the last inequality hold. They establish the uniform asymptotic rate; for a finite parameter choice, use the exact score and gain conditions in (30). If $m \ge 2$, then $\eta x$ is bounded below by a positive absolute constant, so $m \asymp \sqrt{x/\eta}$ and $2m(k-1) = \Omega(\sqrt{HxD})$. If $m = 0$, the score equation gives $x = O(1)$; the factorial boundary supplies $\Omega(D)$ transitions, which is the required rate. Bounded $D \ge 32$ is handled as above by one transition and bounded target rate. Since $U = \Theta(D)$, the upper bound (9) matches this count. In particular, $x = D$ and $H \asymp D$ give $\Theta(D^{3/2})$.

**An exact localization budget.** Suppose an $h$-head scalar construction on $[0,1]$ has score budget $x_0$, $T$ transitions at threshold one, head amplitudes at most $A_*$ throughout, and witness slack at least $\eta_* > 0$. Give $J$ copies cores of length $\ell = L/(3J)$ separated by gaps $2\ell$. Rescale each denominator to its core and multiply by $1 + e^{M(a-z)} + e^{M(z-b)}$, with $a, b$ at distance $\ell/2$ outside it. Set $M\ell = 2\max\{x_0, \log(2(1+\delta)h(J+1)A_*/\eta_*)\}$. On its own core each amplitude changes by at most $2A_* e^{-M\ell/2}$. At distance $d \ge 2\ell$ from another core, the ungated amplitude is at most $A_* e^{x_0 d/\ell}$; the gate is at least $e^{M(d-\ell/2)}$. Thus total gain-weighted error is at most $(1+\delta)h(J+1)A_* e^{-M\ell/2} \le \eta_*/2$. Slopes have magnitude at most $x_0/\ell + M$, giving

$$x_{\text{new}} \le 3J\left[x_0 + 2\max\left\{x_0, \log\frac{2(1+\delta)h(J+1)A_*}{\eta_*}\right\}\right]. \tag{31}$$

Positive constant atoms survive multiplication. Removing repeated boundary labels retains $JT$ transitions with $hJ$ heads. In particular, a Gaussian core with $h = k$, $T = 2m(k-1)$, and

$$\gamma = 2\delta\Theta_\eta e^{\eta(m^2+1)/4} P_k(a_{\eta,p}) S_k(a_{\eta,p}) < 1$$

has $A_* \le \gamma/\delta$, $\eta_* \ge 1 - \gamma$. For integers $J \ge 2$ and $Jk \le H$, the explicit budget

$$3J\left[x_0 + 2\max\left\{x_0, \log\frac{2(1+\delta)k(J+1)\gamma}{\delta(1-\gamma)}\right\}\right] \le x$$

guarantees $2Jm(k-1)$ transitions. For $J = 1$, no gating is needed and the budget is $x_0 \le x$. These integer guarantees can be maximized jointly with $K_C$ and $K_G$; the following argument establishes their uniform growth rate.

For $D \le x \le HD$, choose $J = \max\{1, \lfloor cx/D \rfloor\}$ with sufficiently small absolute $c > 0$, and $k = \lfloor H/J \rfloor$. If $J = 1$, then $x = O(D)$ and the score-$D$ construction already gives $\Omega(\sqrt{HxD})$. Otherwise $J \asymp x/D$, $k \asymp H/J \ge 2$, and each core gives $\Omega(D\sqrt{k})$ transitions. Its amplitude bound $A_* \le \gamma/\delta$ and slack $\eta_* \ge 1 - \gamma$ give localization cost $O(JD)$, so decreasing $c$ makes it fit $x$. The total count is $\Omega(JD\sqrt{k}) = \Omega(\sqrt{HxD})$. Pad unused slots with zero-readout heads. With Appendix C, this proves Theorem 5 throughout its region.

## E Proof of the Two-Head Saturation Construction

For Theorem 6, construct two ideal reciprocal responses with strict gain margins, then approximate their denominators by finite positive exponential sums.

**Parameters and ideal denominators.** Put $t_0 = 2\delta^{-1/3}$, $a = 2D/3 - \log 128$, $\alpha = a/G$, and $\beta = t_0\Gamma(\alpha)/2$. For $D \ge 20$ and $G \ge D$, $a > 0$, $a = \Theta(D)$, and $0 < \alpha < 2/3$. Choose $\omega > 0$ so

that $\frac{|\Gamma(\alpha+i\omega)|}{\Gamma(\alpha)} = 16\delta$. With $\phi = \arg\Gamma(\alpha + i\omega)$, let

$$A(z) = \Gamma(\alpha)z^{-\alpha},$$
$$f_0(z) = \frac{1}{2}\int_0^\infty e^{-pz}p^{\alpha-1}\cos(\omega\log p - \phi)\,dp = 8\delta A(z)\cos(\omega\log z),$$
$$F(z) = \frac{A(z)^2}{2\beta} - 4\delta A(z) + f_0(z), \qquad Q_\pm(z) = A(z) \pm F(z).$$

For $1 \le z \le e^G$, define

$$t(z) = \frac{2\beta}{A(z)} = t_0 z^\alpha, \qquad h(z) = \frac{F(z)}{A(z)} = \frac{1}{t(z)} + \delta\{8\cos(\omega\log z) - 4\}.$$

The parameter choice gives

$$2\delta^{-1/3} \le t(z) \le \frac{1}{64\delta}, \qquad |h(z)| \le \frac{1}{2}\delta^{1/3} + 12\delta \le \delta^{1/3}. \tag{32}$$

Thus $Q_\pm > 0$. With readouts $+\beta$ for $Q_-$ and $-\beta$ for $Q_+$, the nominal response and total absolute amplitude are

$$R(z) = \frac{\beta}{Q_-(z)} - \frac{\beta}{Q_+(z)} = \frac{1 + \delta\{8\cos(\omega\log z) - 4\}t(z)}{1 - h(z)^2},$$
$$T(z) = \frac{\beta}{Q_-(z)} + \frac{\beta}{Q_+(z)} = \frac{t(z)}{1 - h(z)^2}. \tag{33}$$

Thus independent head gains give exactly the interval $[R - \delta T, R + \delta T]$.

**Robust witnesses.** At every cosine zero, (32) implies $h^2 \le \delta^{2/3} \le \delta t/2$, and hence

$$1 - (R + \delta T) = \frac{3\delta t - h^2}{1 - h^2} \ge \frac{5}{2}\delta t > 0, \qquad R - \delta T = \frac{1 - 5\delta t}{1 - h^2} > 0.$$

These witnesses are robustly inactive at absolute threshold one. At a positive cosine maximum, $R - \delta T - 1 = \frac{3\delta t + h^2}{1 - h^2} \ge 3\delta t > 0$, giving robust activity. Selecting one zero between successive positive maxima gives

$$K \ge 2\left\lfloor \frac{\omega G}{2\pi} \right\rfloor \ge \frac{\omega G}{\pi} - 2. \tag{34}$$

Cosine minima need not be witnesses.

**Frequency estimate.** Write $k = \omega G$. Euler's gamma product gives

$$\frac{|\Gamma(\alpha + i\omega)|^2}{\Gamma(\alpha)^2} = \prod_{n=0}^{\infty}\left(1 + \frac{\omega^2}{(n+\alpha)^2}\right)^{-1}.$$

The ratio decreases continuously from one to zero, ensuring existence. Isolate $n = 0$, use $n + \alpha \ge n$ for $n \ge 1$, and apply $\prod_{n\ge1}(1+\omega^2/n^2) = \sinh(\pi\omega)/(\pi\omega)$ gives $\log\frac{\Gamma(\alpha)}{|\Gamma(\alpha+i\omega)|} \le \log(1+\omega/\alpha) + \frac{\pi\omega}{2}$. Consequently

$$D - \log 16 = \log\frac{\Gamma(\alpha)}{|\Gamma(\alpha + i\omega)|} \le \log\left(1 + \frac{k}{a}\right) + \frac{\pi k}{2G}.$$

Thus $k \ge k_*$, where $k_* > 0$ is the unique root of $f(k_*) := \log(1 + k_*/a) + \pi k_*/(2G) = D - \log 16$. For the uniform rate, put $v = \log(D/(\delta G)) \in [0, D]$, $C_a = 20/(40/3 - \log 128)$, and $k_0 = cG(1+v)$. For $\lambda = \pi c/2 < 1$,

$$e^{f(k_0)-D} \le e^{-D+\lambda(1+v)} + cC_a(1+v)e^{-v+\lambda(1+v)} \le e^{-20(1-\lambda)+\lambda} + \frac{cC_a}{1-\lambda}.$$

The last bound uses the maximum $(1-\lambda)^{-1}$ of the second exponential factor. For sufficiently small absolute $c > 0$, this is less than $1/16$, proving $k_* = \Omega(G(1+v))$ uniformly. Also $f(2\pi) \le \log(1 + 2\pi/(40/3 - \log 128)) + \pi^2/20 < 20 - \log 16$, so $k_* > 2\pi$. Consequently $2\lfloor k_*/(2\pi)\rfloor \ge k_*/(2\pi)$. This proves the exact guarantee and its asymptotic rate in (17).

**Positive finite denominators.** The square-envelope identity

$$A(z)^2 = \frac{\Gamma(\alpha)^2}{\Gamma(2\alpha)} \int_0^\infty e^{-pz} p^{2\alpha-1}\, dp$$

expresses $Q_\pm$ as Laplace transforms of the signed densities

$$q_\pm(p) = p^{\alpha-1}\left[1 \pm \left\{\frac{\Gamma(\alpha)}{t_0\Gamma(2\alpha)} p^\alpha - 4\delta + \frac{1}{2}\cos(\omega \log p - \phi)\right\}\right]. \tag{35}$$

Set $B = 8D$, $p_0 = e^{-3D/\alpha}$, and $C_\alpha = \Gamma(\alpha)/(t_0\Gamma(2\alpha))$. Log convexity gives $\Gamma(1+\alpha) \le 1$, and integration over $[1,2]$ gives $\Gamma(1+u) \ge g_0 := e^{-1} - e^{-2}$ for $0 \le u \le 4/3$. Thus $C_\alpha \le \delta^{1/3}/g_0$. The density bound

$$b_q := C_\alpha B^\alpha + 4\delta + \tfrac{1}{2} < 1 \tag{36}$$

holds throughout $D \ge 20$: its upper bound $e^{-D/3}(8D)^{2/3}/g_0 + 4e^{-D} + 1/2$ is below one at $D = 20$ and decreases thereafter. In particular, both densities are positive on $[p_0, B]$ and at most $(1+b_q)p^{\alpha-1}$.

The uniform absolute truncation error of either denominator is bounded by

$$E_{\text{tail}} = (\tfrac{3}{2} + 4\delta)\left[\frac{p_0^\alpha}{\alpha} + \Gamma(\alpha, B)\right] + C_\alpha\left[\frac{p_0^{2\alpha}}{2\alpha} + \Gamma(2\alpha, B)\right],$$

where $\Gamma(u, B) = \int_B^\infty p^{u-1}e^{-p}\, dp$. Here $A_{\min} = 128\Gamma(\alpha)\delta^{2/3}$. Using $\Gamma(\alpha, B) \le e^{-B}$, $\Gamma(2\alpha, B) \le 2e^{-B/2}$, and $\alpha\Gamma(\alpha) \ge g_0$ gives $E_{\text{tail}}/A_{\min} \le \delta^{7/3}$ for $D \ge 20$.

**A margin-driven context budget.** Put $b = \delta^{1/3}$ and define the exact relative error allowance

$$e_* := \frac{\delta(1-b)^2}{1+\delta+\delta(1-b)}, \qquad \Delta = e_* - E_{\text{tail}}/A_{\min} > 0.$$

Choose the largest $h \in (0, 1]$ satisfying

$$C_{\text{quad}}(h) := \frac{(1+b_q)h^2}{8}\alpha(\alpha+1)e^{h(\alpha+4)} \le \Delta/2,$$

and put $M = \lceil \log(B/p_0)/h \rceil$, $s = \log(B/p_0)/M \le h$. Use nodes $p_j = p_0 e^{js}$ and positive endpoint weights

$$w^L_{\pm,j} = \int_{p_j}^{p_{j+1}} \frac{p_{j+1} - p}{p_{j+1} - p_j} q_\pm(p)\, dp, \qquad w^R_{\pm,j} = \int_{p_j}^{p_{j+1}} \frac{p - p_j}{p_{j+1} - p_j} q_\pm(p)\, dp.$$

These weights have elementary antiderivatives. Combining endpoints gives $\nu_{\pm,j} > 0$ and quadrature exact for each bin's mass and first moment. Linear interpolation of $e^{-pz}$ on a bin has error between zero and $z^2(p_{j+1} - p_j)^2 e^{-p_j z}/8$. Since $p_{j+1} - p_j \le he^h p$ and $p_j \ge e^{-h}p$, integration bounds the quadrature error by $C_{\text{quad}}(h)A(z)$. Add source and query atoms of weight $\zeta = A_{\min}\Delta/4$ each. Then $|\widetilde{Q}_\pm - Q_\pm| \le e_* A$ throughout the path.

Since $Q_\pm \ge (1-b)A$, either gain-response endpoint changes by at most

$$\frac{(1+\delta)te_*}{(1-b)(1-b-e_*)} = \delta t.$$

This preserves the strict witness margins from (33)–(34). Source readouts $\pm\beta/\zeta$ and background logits $-p_j z + \log(\nu_{\pm,j}/\zeta)$ realize these finite sums. Since $\Delta = \Theta(\delta)$ uniformly, $h = \Omega(\sqrt{\delta})$. The context size is therefore

$$M + 3 = \left\lceil \frac{\log(8D) + 3DG/a}{h} \right\rceil + 3 = O(G/\sqrt{\delta}). \tag{37}$$

All asymptotic constants are absolute and uniform over the theorem's domain. □

# F Fixed-Head Euclidean Upper Bound at High Score

Fix $H_0 \geq 2$. We prove that, for $2 \leq H \leq H_0$, $D = \log(1/\delta) > 19$, $x = BL$, and $D \leq \bar{G} := 1 + \log(1 + x/D) \leq D/\delta$, the Euclidean robust transition count satisfies

$$K \leq C_{H_0} \bar{G} \left[ 1 + \log\left( 1 + \frac{D}{\delta \bar{G}} \right) \right]. \tag{38}$$

Constants may depend on $H_0$, but not context length or term count.

**Reduction to a fixed set of heads.** At threshold one, write $y(z) = \sum_{h=1}^{H} b_h / Q_h(z)$ with positive exponential sums $Q_h$ supported in $[-B, B]$. Set $a_h(z) = \|b_h\|_2 / Q_h(z)$, ignoring zero-readout heads. At an inactive witness, the gain-box support bound (22) gives $a_h(z) \leq 1/(2\delta)$. Concavity of $\log a_h$ makes $\{z : a_h(z) > 1/(2\delta)\}$ an interval, containing no inactive and at most one active witness. Delete these active witnesses, costing at most two transitions per head. Of at most $H + 1$ remaining runs, one has at least $\max\{0, \lceil (K - 2H)/(H + 1) \rceil\}$ transitions.

Fix $0 < c < 1/4$ and put $\tau = c\delta/H$. The intervals $\{a_h \geq \tau\}$ give at most $2H$ membership changes. Split there and retain a longest subrun with $m$ transitions, so $K \leq 2H(H+2) + (H+1)(2H+1)m$, with a fixed significant-head set $S$. At its witnesses,

$$\tau \leq a_h \leq 1/(2\delta) \quad (h \in S), \qquad \sum_{h \notin S} a_h \leq c\delta. \tag{39}$$

Give $S$ gains in $[1 - \delta/2, 1 + \delta/2]^S$, others gain one. Scaling by $(1 - \delta)/(1 - \delta/2)$ puts them in the original box; the unscaled norm exceeds $(1 - \delta/2)/(1 - \delta)$ at active witnesses and is at most one at inactive witnesses. Dropping heads outside $S$ costs at most $c\delta$. The reduced half-box responses therefore alternate about $\theta_\delta = 1 + \delta/[4(1 - \delta)]$ with margin at least $\mu_\delta = \delta[1/(4(1 - \delta)) - c] > 0$.

For $h \in S$, the witness range of $\log Q_h$ has length at most

$$R_0 = \log(1/(2\delta\tau)) = \log(H/(2c\delta^2)) = O_H(D). \tag{40}$$

By convexity of $\log Q_h$, split at each minimum of $\log Q_h$ that breaks monotonicity, cutting its gap or assigning a minimizing witness once. This costs at most $H$ transitions. On each branch, every $\log Q_h$ is monotone with endpoint variation at most $R_0$.

**A local reciprocal approximation.** Let $Q(x) = \sum_\nu c_\nu e^{\lambda_\nu x} > 0$, and let $J$ be an interval of length $s$, centered at $x_0$, such that $\log Q$ is monotone on $J^* = [x_0 - 4s, x_0 + 4s]$. Write $d_J = \operatorname{osc}_{J^*} \log Q$ and set $d_0 = \log(37/21)$, $q_J = 21(e^{d_J} - 1)/64$. For $0 < d_J \leq d_0$ and $0 < \eta \leq 1/4$, take

$$n = \max\left\{ 0, \left\lceil \log_2 \frac{e^{2d_J} q_J}{(1 - q_J)\eta} \right\rceil \right\}.$$

Then a polynomial $p_n$ of degree at most $n$ satisfies

$$\left| e^{-\mu x} p_n(x) - \frac{1}{Q(x)} \right| \leq \frac{\eta}{Q(x)}, \qquad \mu = \frac{\log Q(x_0 + 4s) - \log Q(x_0 - 4s)}{8s}. \tag{41}$$

For $d_J = 0$, take $n = 0$.

Put $E(x) = e^{-\mu x} Q(x)$ and normalize its weights at $x_0$ to a probability law for $u = \lambda - \mu$. Equal endpoint values give $\mathbb{E} \sinh(4su) = 0$ and $\mathbb{E} \cosh(4su) \leq e^{d_J}$. For complex $|\zeta| \leq 1/4$, the even terms of $e^{\zeta w} - 1 - \zeta \sinh w$ are bounded by $(\cosh w - 1)/16$; its odd terms are bounded by $17(\sinh|w| - |w|)/64 \leq 17(\cosh w - 1)/64$. Thus $|E(x_0 + 4s\zeta)/E(x_0) - 1| \leq q_J \leq 1/4$. The analytic function $E(x_0)/E(x) - 1$ is bounded by $q_J/(1 - q_J)$ on the radius-$s$ disk. Its degree-$n$ Taylor tail on $J$ is at most $[q_J/(1 - q_J)]2^{-n}$. Multiplying by $E(x)/E(x_0) \leq e^{2d_J}$ gives the stated relative error and the explicit degree.

**Local zero counts and exceptional cells.** On a small-deficit interval $J$ where every $d_{h,J} \leq d_0$, apply (41) to all $h \in S$ with $\eta = \delta/2$. The reciprocal replacements give gains in $[1 - \delta/2, 1 + \delta/2]$, within the half box, preserving every strict label about $\theta_\delta$. The squared norm minus $\theta_\delta^2$ is an exponential polynomial with at most $t_S = |S|(|S| + 1)/2 + 1$ exponents and polynomial coefficients

of degree at most $2n_J$, where $n_J$ is the largest approximation degree on $J$. The zero count is at most the sum of polynomial degrees plus the number of exponents minus one: remove one exponential factor, differentiate one more time than its polynomial degree, then use Rolle's theorem and induction on the exponent count. Thus $J$ contributes at most

$$t_S(2n_J + 1) - 1, \qquad d_J := \sum_{h \in S} d_{h,J}. \tag{42}$$

For a large-deficit interval, use the exact cleared expression. Put $Q_S = \prod_{h \in S} Q_h$, $N_S = \sum_{h \in S} b_h Q_S / Q_h$, and

$$F_S = \|N_S\|_2^2 - (\theta_\delta)^2 Q_S^2, \qquad A_S = (\theta_\delta)^2 Q_S^2 + \left( \sum_{h \in S} \|b_h\|_2 Q_S / Q_h \right)^2.$$

The envelope $A_S$ dominates $F_S$'s absolute coefficients. The half-box margin and (39) give the finite relative margin

$$|F_S| \geq \rho A_S, \qquad \rho = \frac{2\theta_\delta \mu_\delta - \mu_\delta^2}{\theta_\delta^2 + H^2/(4\delta^2)} = \Theta_H(\delta^3).$$

For $r + 1$ alternating witnesses in $J$, let $\mu$ be the secant slope of $\log A_S$ across $J^*$, and set $E = e^{-\mu x} A_S$. Write $d_{A,J} = \operatorname{osc}_{J^*} \log A_S$. The order-$r$ divided difference of $e^{-\mu x} F_S$ has magnitude at least $(r + 1)\rho \min_J E / s^r$, since its summands have a common sign and each denominator is at most $s^r$. The mean-value theorem and $|u|^r \leq 2r!(3s)^{-r}\{\cosh(3su) - 1\}, \qquad r \geq 2$, bound the same derivative above. For even $r \geq 2$, the cosine-hyperbolic series gives $|w|^r \leq r!(\cosh w - 1)$. For odd $r \geq 3$, taking the geometric mean of the neighboring even-power bounds gives the factor $\sqrt{(r-1)!(r+1)!} = r!\sqrt{(r+1)/r} \leq 2r!$; set $w = 3su$. The secant tilt gives $\operatorname{osc}_{J^*} \log E \leq 2d_{A,J}$, hence $(r + 1)\rho 3^r \leq 2e^{2d_{A,J}}\left(e^{2d_{A,J}} - 1\right)$. Each positive term of $A_S$ contains at most two copies of each $Q_h$, so $d_{A,J} \leq 2d_J$. The finite count for a large-deficit cell is

$$R_J = \max\left(\{1\} \cup \{r \in \mathbb{N},\ r \geq 2 : (r + 1)\rho 3^r \leq 2e^{4d_J}(e^{4d_J} - 1)\}\right).$$

This implies $R_J = O_H(D + d_J)$; the integer maximum is the bound used before asymptotic summation.

**Global partition and summation.** On a monotone branch, put $\ell_0 = D/B$. If the branch length is at most $2\ell_0$, treat it as one boundary core. Otherwise, for a branch $[a, b]$ with midpoint $c$, keep end cores of length $\ell_0$. Divide each nonempty layer $[a + 2^j \ell_0, \min\{a + 2^{j+1}\ell_0, c\}]$ into eight equal cells, and reflect this construction about $c$ for the right half. Each cell's enlargement $J^*$ stays inside $[a, b]$. The enlargements have bounded overlap: at endpoint distance $2^j \ell_0$, cell lengths are at most $2^j \ell_0 / 8$, so a point meets only a fixed number of neighboring layers. There are at most $M \leq 2 + 16\lceil \log_2 \max\{1, BL/(2D)\} \rceil$ cells, including the cores. For interior cells, monotonicity and (40) give

$$V := \sum_J d_J \leq \nu |S| R_0, \qquad \nu := \max_z \#\{J : z \in J^*\} \leq 54, \tag{43}$$

by integrating $\sum_{h \in S} |(\log Q_h)'|$ over the enlargements. Within one layer a point lies in at most nine enlarged cells; each half-branch contributes at most three layers, giving $\nu \leq 54$. Assign shared endpoints once. With $d_J > d_0$, there are at most $\lfloor V/d_0 \rfloor$ large-deficit intervals, where $V = \sum_J d_J$; their total finite count is $\sum_{J \text{ large}} R_J$. The variation bound implies this total is $O_H(D^2)$.

Every end core $I$ has length $|I| \leq 2D/B$. Its spectral bound is

$$B_I = \left\lfloor \min_{n \in \mathbb{N}} \left[ n - 1 + \frac{n \log(1/\rho) + 48HB|I|/n}{\log 2} \right] \right\rfloor.$$

On a branch with $M$ cells, the finite transition bound is

$$T_{\text{branch}} \leq M - 1 + \sum_{I \text{ core}} B_I + \sum_{J \text{ small}} [t_S(2n_J + 1) - 1] + \sum_{J \text{ large}} R_J.$$

There are at most $H+1$ monotonicity branches and at most $H$ connecting transitions. Undoing the reductions gives the finite bound

$$K \le 2H(H+2) + (H+1)(2H+1)[H + (H+1)\max T_{\text{branch}}].$$

Thus all cell counts and reduction losses are explicit; only the following growth-rate conclusion uses asymptotic notation.

For the small-deficit intervals, sum (42), use (43), and apply Jensen's inequality:

$$\sum_{J \text{ small}} \left[1 + \log\left(1 + \frac{d_J}{\delta}\right)\right] \le C_H \bar{G}\left[1 + \log\left(1 + \frac{D}{\delta \bar{G}}\right)\right].$$

Finally, $t[1 + \log(1 + D/(\delta t))]$ is increasing in $t > 0$, so $\bar{G} \ge D$ gives

$$\bar{G}\left[1 + \log\left(1 + \frac{D}{\delta \bar{G}}\right)\right] \ge D[1 + \log(1 + 1/\delta)] \ge D^2.$$

Thus the boundary and exceptional-cell costs are absorbed into (38). The finite reduction formula above proves the claim.

For the parameterization of Theorem 6, $x = 8D(e^G - 1)$ implies $G \le \bar{G} \le G + 1 + \log 8$. If $\bar{G} \le D/\delta$, (38) is $O_H(G[1 + \log(D/(\delta G))])$. Otherwise $G > D/\delta - 1 - \log 8 = \Theta(D/\delta)$, and the unrestricted upper bound gives the same estimate. The two-head lower construction, padded by zero-readout heads and rescaled to any $B, L$ with this product, proves (18). □

## G THE VANISHING-SCORE BOUNDARY

For (11)–(12), write $x = BL$, $D = \log(1/\delta)$, and $\tau_\delta = \log((1+\delta)/(1-\delta))$.

**Gain-box overlap.** Since each source-relative score slope has magnitude at most $B$, $|\partial_z \log w_h| \le B$. If $B|p-q| \le \tau_\delta$, the intervals $[(1-\delta)w_h(p), (1+\delta)w_h(p)]$ and $[(1-\delta)w_h(q), (1+\delta)w_h(q)]$ intersect headwise. Common effective weights give admissible gains at $p$ and $q$ with identical outputs, forbidding opposite robust labels. Each transition requires $B|p-q| > \tau_\delta$; summing gives $K < x/\tau_\delta$. When $B = 0$, all source weights are constant and $K = 0$.

Conversely, on $[0, L]$ one effective head can have source weight $w(z) = (2 + Me^{-Bz})^{-1}$. Its endpoint ratio tends to $e^x$ as $M \to \infty$. If $x > \tau_\delta$, take finite $M$ so that $(1-\delta)w(L) > (1+\delta)w(0)$ and place the threshold strictly between them. This gives one finite-attention transition, with zero readout on the other heads, proving the exact zero-capacity assertion for every norm.

For one head, the exact score-dependent capacity is

$$\mathcal{C}^{\text{Euc}}_{B,L}(1,\delta) = \begin{cases} 0, & x \le \tau_\delta, \\ 1, & \tau_\delta < x \le 2\,\text{arcosh}(e^{\tau_\delta}), \\ 2, & x > 2\,\text{arcosh}(e^{\tau_\delta}). \end{cases} \tag{44}$$

Unimodality permits at most two transitions; one exists precisely when $x > \tau_\delta$. Two require inactive–active–inactive at $q_- < p < q_+$. Write $Q = 1/w$ as a positive exponential sum with source-relative exponents in $[-B, B]$, and let $R$ be its exponent under the normalized weights at $p$. With $a = p - q_-$, $b = q_+ - p$, and $\mu = \mathbb{E}R/B$, convexity gives

$$\frac{Q(q_-)}{Q(p)} \le \cosh(Ba) - \mu \sinh(Ba), \qquad \frac{Q(q_+)}{Q(p)} \le \cosh(Bb) + \mu \sinh(Bb).$$

The maximum of the smaller bound occurs at $\mu = \tanh(B(a-b)/2)$, giving $\cosh(B(a+b)/2)/\cosh(B(a-b)/2) \le \cosh(x/2)$. Both endpoint ratios must exceed $e^{\tau_\delta}$, proving necessity. Conversely, $Q(z) = 2 + M(e^{Bz} + e^{-Bz})$ on $[-L/2, L/2]$ has endpoint-to-center ratio tending to $\cosh(x/2)$ as $M \to \infty$; finite $M$ and a threshold between the robust levels realize two transitions under the strict inequality. Only positive scalar weights enter the argument, so these thresholds hold for every output norm.

For $H \ge 2$, the two-transition onset is also exact: $\mathcal{C}^{\text{Euc}}_{B,L}(H,\delta) \ge 2$ if and only if $x > 2\tau_\delta$. Necessity follows from $K < x/\tau_\delta$. For sufficiency, use the limiting scalar contributions $ce^{Bz}$ and $-ce^{-Bz}$ on

$[-L/2, L/2]$. Their uncertainty radius at zero is $2c\delta$, while at either endpoint the smallest perturbed magnitude is $2c[\sinh(x/2) - \delta\cosh(x/2)]$. A separating threshold exists exactly when $x > 2\tau_\delta$. Replacing the contributions by $cM/(2+Me^{-Bz})$ and $-cM/(2+Me^{Bz})$ preserves the strict margins for large finite $M$ and uses finite standard attention.

**A polynomial upper bound.** Suppose $0 < x \le 1/4$ and $n \ge 0$ is an integer such that

$$\inf_{x/2<u<\log 2} \frac{e^{x/2}}{2-e^u}\left(\frac{x}{2u}\right)^{n+1} \le \delta. \tag{45}$$

Then Euclidean capacity is at most $2n$, independently of $H$ and output dimension. Center at $z_c$, set $t = B(z - z_c)$, and normalize each denominator:

$$q_h(t) = \frac{Q_h(z_c + t/B)}{Q_h(z_c)} = \sum_j p_j e^{r_j t}, \qquad p_j > 0, \quad \sum_j p_j = 1, \quad |r_j| \le 1.$$

The absolute Taylor coefficients of $q_h - 1$ are majorized by those of $e^t - 1$. The geometric expansion $1/q_h = \sum_{k\ge 0} a_{hk} t^k$ has majorant $(2 - e^t)^{-1}$, giving $\sum_k |a_{hk}|u^k \le (2-e^u)^{-1}$ for $0 < u < \log 2$. Since $|t| \le x/2$, the degree-$n$ truncation $p_h(z)$ satisfies

$$\left|\frac{p_h(z)}{w_h(z)} - 1\right| \le \frac{e^{x/2}}{2-e^u}\left(\frac{x}{2u}\right)^{n+1}, \qquad x/2 < u < \log 2.$$

The infimum is attained in this open interval whenever it is at most $\delta < 1$: the left endpoint gives a bound greater than one and the right endpoint diverges. Hence (45) provides an admissible relative-$\delta$ approximation. Thus $P(z) = \sum_h p_h(z) B_h$ is an admissible point-dependent gain response. The polynomial $\|P(z)\|_2^2 - \varepsilon^2$, of degree at most $2n$, alternates signs at the witnesses. Slightly increasing $\varepsilon^2$ makes inactive inequalities strict while preserving activity, so Rolle's theorem gives $K \le 2n$.

Thus the first critical interval is exact. For $H \ge 2$, $0 < \delta \le 1/256$, and $0 \le x \le \sqrt{\delta}/2$,

$$\mathcal{C}^{\mathrm{Euc}}_{B,L}(H,\delta) = \begin{cases} 0, & x \le \tau_\delta, \\ 1, & \tau_\delta < x \le 2\tau_\delta, \\ 2, & 2\tau_\delta < x \le \sqrt{\delta}/2. \end{cases} \tag{46}$$

Indeed, overlap gives $K < x/\tau_\delta$, the preceding polynomial bound with $n = 1$ and $u = 1/2$ gives $K \le 2$ (the bound is less than $4x^2 \le \delta$), and the exact onset above supplies both lower bounds.

**Uniform inversion for all head counts.** Let $D \ge 64$, $H \ge 2$, and $\tau_\delta < x \le 1$, and put $A_0 = \log\big(e + (1 + D/H)/x\big), \qquad q = D/A_0$. Since $x > \tau_\delta > 2\delta$ and $1 + D/H \le D$, $A_0 \le D + \log D \le 2D$, so $q \ge 1/2$.

For the upper bound, first suppose $H \le D$. The factorial inequality in Appendix C is valid for every $x > 0$ and gives

$$K \le 1 + C\left[X + \frac{d}{\log(e + d/X)}\right], \qquad X = 4Hx, \quad d = 1 + \log\frac{(1+\delta)^2(H+\delta^2)}{\delta^3(2+\delta)} = \Theta(D).$$

Here $Hx \le D$ and $A_0$ is comparable to $\log(e + D/(Hx))$. The elementary bound $t\log(e + 1/t) = O(1)$ for $0 < t \le 1$ absorbs $X$ into the second term, giving $K = O(q)$. If $H > D$ and $x \le 1/4$, choose $n = \min\{j \ge 0 : (45) \text{ holds}\}$ in (45). Then $K \le 2n = O(q)$. If $H > D$ and $1/4 < x \le 1$, divide the path into four equal cells and apply the same polynomial bound on each with score product $x/4$. If $n_4$ is the smallest integer satisfying (45) with $x$ replaced by $x/4$, adding at most three connecting transitions gives the finite bound $K \le 8n_4 + 3 = O(D) = O(q)$, with absolute upper constants.

The finite lower bound is $K \ge K_C$, obtained by maximizing $2m(k-1)$ over the explicit set $\mathcal{F}_C$ in Appendix D; if $x > \tau_\delta$, also $K \ge 1$. For its uniform rate, fix $\gamma \in (0,1)$ and take a sufficiently small absolute $c > 0$. When $M = \lfloor cq \rfloor \ge 1$, set $k = \min\{H, M+1\}$ and $m = \lfloor M/(k-1) \rfloor$. Then $m(k-1) = (k-1)\lfloor M/(k-1)\rfloor \in [M/2, M]$ and $m \le 1 + 2cD/H$. Use the transformed

Chebyshev envelope of Appendix D with the full score product $x$. Its exact maximum and dynamic range are

$$E = \cosh\left(m \operatorname{arcosh}\left(3 + \frac{4}{e^{x/m} - 1}\right)\right), \qquad R \le e^x.$$

The finite certificate (27) applies with $a = 1/E$. Since $\log E = O(m \log(e + m/x))$ and $\log(e + m/x) = O(A_0)$, its logarithmic budget excluding $\delta$ is $x + O(MA_0 + 1) = O(cD + 1)$. Choosing $c$ sufficiently small certifies $\Omega(q)$ transitions for all sufficiently large $D$. Bounded $q$, and the remaining bounded-$D$ region $D \ge 64$, have bounded target rate and are covered by the one-transition construction. Thus both constants in (12) are absolute, independent of $H, x, \delta$. Zero-readout padding preserves finite attention.

## H A Two-Head Mellin Bridge Below Saturation

For (15), let $1 \le G \le D$, $\alpha = 1/G$, $\gamma = e^G$, and $r = (4e + 2)\delta$. Since $D \ge 32$, $0 < r < 1/2$. For $\omega > 0$, the Mellin–Laplace identity gives

$$\int_0^\infty e^{-pz} p^{\alpha + i\omega - 1}\, dp = \Gamma(\alpha + i\omega) z^{-\alpha - i\omega}.$$

The ratio $|\Gamma(\alpha + i\omega)|/\Gamma(\alpha)$ decreases continuously from one to zero. Choose $\omega$ with ratio $r$ and $\phi = \arg \Gamma(\alpha + i\omega)$; then

$$A(z) = \Gamma(\alpha) z^{-\alpha}, \qquad f(z) = \int_0^\infty e^{-pz} p^{\alpha - 1} \cos(\omega \log p - \phi)\, dp = rA(z) \cos(\omega \log z).$$

Both $Q_\pm = A \pm f$ are Laplace transforms of the nonnegative densities $p^{\alpha-1}[1 \pm \cos(\omega \log p - \phi)]$. Let $A_{\min} = A(\gamma) = \Gamma(\alpha)/e$ and $\beta = A_{\min}/(8\delta)$. Give the two reciprocal heads coefficients $+\beta$ and $-\beta$. Their nominal response $g$ and total absolute amplitude $T$ satisfy

$$g = \frac{\beta}{Q_+} - \frac{\beta}{Q_-}, \qquad T = \frac{\beta}{Q_+} + \frac{\beta}{Q_-}, \qquad \frac{|g|}{T} = r|\cos(\omega \log z)|.$$

At every cosine zero, $|g| + \delta T \le 1/4$. At each cosine extremum, using $A(1) = eA_{\min}$,

$$|g| - \delta T = \frac{(r - \delta)2\beta}{A(1 - r^2)} \ge \frac{4e + 1}{4e(1 - r^2)} > 1.$$

Thus extrema and intervening zeros alternate robust activity and inactivity at threshold one. With $k = \omega G$, there are at least $2\lfloor k/\pi \rfloor$ transitions.

Euler's gamma product, with the $n = 0$ factor separated, yields

$$D - \log(4e + 2) = \log \frac{\Gamma(\alpha)}{|\Gamma(\alpha + i\omega)|} \le \log(1 + k) + \frac{\pi k}{2G}.$$

Its right side is strictly increasing, so $k \ge k_\dagger$, where $\log(1 + k_\dagger) + \pi k_\dagger/(2G) = D - \log(4e + 2)$. For sufficiently small absolute $c > 0$, evaluation at $cDG$ gives $\log(1 + cD^2) + \pi cD/2 < D - \log(4e + 2)$ uniformly for $D \ge 32$. Indeed, $\log(1 + cD^2)/D$ is uniformly small as $c \downarrow 0$; the remaining term is linear in $c$. Thus $k_\dagger = \Omega(DG)$. Also $k_\dagger > \pi$, since $\log(1 + \pi) + \pi^2/2 < 32 - \log(4e + 2)$. The exact count $2\lfloor k_\dagger/\pi \rfloor$ is consequently $\Omega(DG)$, proving (15).

For finite heads, let $q_* = (1 - r)A_{\min}$ and let $\mu = \min\{3/4, (4e + 1)/(4e(1 - r^2)) - 1\} > 0$ be the witness slack. Choose a uniform denominator error allowance

$$0 < \eta < \frac{\mu q_*^2}{4\beta(1 + \delta) + \mu q_*}.$$

Either gain-response endpoint then changes by less than $\mu/2$, since the reciprocal error is at most $\eta/[q_*(q_* - \eta)]$ per head. One may choose $\eta = \Theta(\delta A_{\min})$ uniformly. Truncate to $[p_0, 8D]$, taking $p_0$ small enough that $2p_0^\alpha/\alpha < \eta/4$. The upper tail is at most $2e^{-8D} < \eta/4$ for an allowance of this order and $D \ge 32$. A common positive quadrature on the compact interval can have error below $\eta/4$; add source and query atoms of total weight below $\eta/4$. The combined error is below $\eta$, preserving every strict label. Each density is nonnegative and not identically zero on any bin, so the quadrature coefficients are positive. The finite denominators have exponents in $[-8D, 0]$ and a positive constant atom, so Appendix A implements them as standard attention on the path $[1, e^G]$ with $B \le 8D$ and $L = e^G - 1$.

# I A FIXED-HEAD LOGARITHMIC-SCORE UPPER BOUND

For (16)'s upper bound, the following counting lemma uses the envelope's witness dynamic range, rather than the spectral bound of Lemma S1.

**Lemma S2** (Bounded-envelope alternation)**.** *Let* $f(z) = \sum_\nu c_\nu e^{\lambda_\nu z}$ *be real, and let* $A(z) = \sum_\nu a_\nu e^{\lambda_\nu z}$ *with* $a_\nu > 0$, $|c_\nu| \le a_\nu$, *and* $|\lambda_\nu| \le \Lambda$. *At* $z_0 < \cdots < z_m$, *suppose* $f$ *alternates signs,* $|f(z_i)| \ge \rho A(z_i)$, *and* $\max_i A(z_i) \le e^R \min_i A(z_i)$ *for* $0 < \rho \le 1$. *Then, with absolute* $C$,

$$m \le C\Gamma \left[1 + \log\left(1 + \frac{\Lambda(z_m - z_0)}{\Gamma}\right)\right], \qquad \Gamma = R + \log(1/\rho) + 1. \tag{47}$$

The finite bound underlying (47) is explicit. Put $Y = \Lambda(z_m - z_0)$ and define

$$F(Y) = \max\{j \in \mathbb{Z}_{\ge 0} : \rho(j+1)! \le e^R Y^j\}, \qquad T_4 = \max\{j \in \mathbb{Z}_{\ge 0} : (j+1)\rho 4^j \le 2e^R\}.$$

For $Y > 0$, both maxima are finite. If $Y \le 4\Gamma$, then $m \le F(Y)$. Otherwise, with $M = 2 + 8\lceil \log_2(Y/(2\Gamma))\rceil$,

$$m \le M - 1 + 2F(\Gamma) + (M-2)T_4.$$

The first formula also covers $m = 0$; $Y = 0$ permits no transitions.

*Proof.* Write $L_0 = z_m - z_0$ and $A_0 = \min_i A(z_i)$. Since $A$ is convex, $A(z) \le e^R A_0$ throughout $[z_0, z_m]$. First suppose $\Lambda L_0 \le 4\Gamma$. Alternating signs make all summands in the order-$m$ divided difference of $f$ agree in sign. Its magnitude is at least $(m+1)\rho A_0 / L_0^m$. The generalized mean-value theorem and $|f^{(m)}| \le \Lambda^m e^R A_0$ imply

$$\rho(m+1)m! \le e^R(\Lambda L_0)^m. \tag{48}$$

Stirling's lower bound $m! \ge (m/e)^m$ gives $m \le C\Gamma$.

A consecutive alternating block of span $\ell$, at distance at least $g > \ell$ from both global endpoint witnesses, has $q$ internal transitions satisfying

$$(q+1)\rho(g/\ell)^q \le 2e^R. \tag{49}$$

Its divided difference has magnitude at least $(q+1)\rho A_0/\ell^q$. At the mean-value point $\xi$, use $u^q \le (q/(eg))^q e^{gu}$ separately for positive and negative exponents. The endpoint guards then give $|f^{(q)}(\xi)| \le 2e^R A_0 (q/(eg))^q$. Combining this with $q! \ge (q/e)^q$ proves (49).

For $\Lambda L_0 > 4\Gamma$, set $h = \Gamma/\Lambda$ and keep boundary cores of length $h$. Divide each half of the remaining interval into endpoint-distance shells $[2^j h, 2^{j+1}h]$, each with four cells, truncating the last shell at the midpoint. This gives at most $M = 2 + 8\lceil \log_2(\Lambda L_0/(2\Gamma))\rceil$ cells. Non-core cells have length at most one quarter of either endpoint distance, so $g/\ell \ge 4$ and (49) gives $T_4$ internal transitions. In boundary cores, (48) with $\Lambda\ell \le \Gamma$ gives the bound $F(\Gamma)$. Adding at most $M-1$ connecting transitions gives the finite formula. Stirling inversion gives $F(\Gamma) = O(\Gamma)$, and taking logarithms gives $T_4 = O(\Gamma)$, yielding (47). □

**Application to Euclidean attention.** Fix $H_0 \ge 2$, $2 \le H \le H_0$, and $0 < \delta \le 1/8$. Ignore heads with zero readout, normalize the threshold to one, and write $y(z) = \sum_h b_h/Q_h(z)$ and $a_h(z) = \|b_h\|_2/Q_h(z)$. At every inactive witness, the gain-box support bound (22) gives $a_h \le 1/(2\delta)$. Concavity of $\log a_h$ makes its superlevel set above $1/(2\delta)$ an interval containing at most one active witness bracketed by inactive witnesses. Delete at most two endpoint active witnesses and at most $H$ overshooting witnesses. Of at most $H+1$ runs, one retains a run with at least $\max\{0, \lceil (K - 2H - 4)/(H+1)\rceil\}$ transitions, with $a_h \le 1/(2\delta)$ headwise.

Fix $0 < c < 1/4$ and put $\tau = c\delta/H$; mark head $h$ significant where $a_h \ge \tau$. These intervals give at most $2H$ membership changes. Split there and retain a longest subrun with fixed significant set $S$, $m$ transitions, and $K \le 2H + 4 + (H+1)[2H + (2H+1)m]$. At its witnesses, $\tau \le a_h \le 1/(2\delta) \quad (h \in S), \qquad \sum_{h \notin S} a_h \le c\delta$. Common gains show that the nominal norm is greater than $(1-\delta)^{-1}$ at active witnesses and at most $(1+\delta)^{-1}$ at inactive witnesses. Removing the insignificant heads changes it by at most $c\delta$. Hence the reduced response $y_S = \sum_{h \in S} b_h/Q_h$ satisfies

$$\left|\|y_S\|_2^2 - 1\right| \ge \mu_c(\delta) := \min\left\{[(1-\delta)^{-1} - c\delta]^2 - 1, 1 - [(1+\delta)^{-1} + c\delta]^2\right\} = \Theta(\delta) \tag{50}$$

with alternating signs.

Clear only denominators in $S$: $Q_S = \prod_{h\in S} Q_h$, $N_S = \sum_{h\in S} b_h Q_S/Q_h$, $F_S = \|N_S\|_2^2 - Q_S^2$, and take the positive coefficient envelope $A_S = Q_S^2 + \left(\sum_{h\in S} \|b_h\|_2 Q_S/Q_h\right)^2$. This dominates the absolute uncollected coefficients of $F_S$. At the witnesses, $|F_S| \geq \mu_c(\delta) Q_S^2$, $Q_S^2 \leq A_S \leq (1 + H^2/(4\delta^2))Q_S^2$. Thus $|F_S| \geq \rho A_S$ for $\rho = \mu_c(\delta)/(1 + H^2/(4\delta^2))$. Also, for $h \in S$, the ratio of $Q_h$ at any two witnesses is at most $H/(2c\delta^2)$. Therefore

$$\max_{i,j} \log \frac{A_S(z_i)}{A_S(z_j)} \leq 2H \log(H/(2c\delta^2)) + \log(1 + H^2/(4\delta^2)) = O_{H_0}(D).$$

The exponents of $F_S$ and $A_S$ have absolute value at most $2HB$. Lemma S2 applies with $\Lambda = 2HB$ and $\Gamma = O_{H_0}(D)$, giving

$$K \leq C_{H_0} D \left[1 + \log\left(1 + \frac{BL}{D}\right)\right]. \tag{51}$$

For $BL = 8D(e^G - 1)$ and $G \geq 1$, the bracket is $O(G)$. Combining this with (15), and padding the two-head construction by zero-readout heads, proves (16).

## J Proof of the Fixed-Norm Separation

Prove Proposition 7 using the norm of Section 7.

*Proof.* Consider the ideal response $y(z) = (e^z, e^{2z})$ on $0 \leq z \leq \log 2$, and write $q = e^z \in [1, 2]$. For every $a \in \mathcal{A}$, $\ell_a(q, q^2) := \frac{2q}{a} - \frac{q^2}{a^2} = 1 - (q/a - 1)^2 \in [0, 1]$. In particular, $\|y(\log a)\|_* = 1$ at each $a \in \mathcal{A}$. Put $a_n = 1 + c_0/\log\log(n + e)$, $\Delta_n = a_n - a_{n+1}$, and $q_n = (a_n + a_{n+1})/2$. No point of $\mathcal{A}$ lies strictly between $a_{n+1}$ and $a_n$. The nearest support point in relative distance is $a_n$, so

$$g_n := 1 - \|y(\log q_n)\|_* = \frac{\Delta_n^2}{4a_n^2}, \qquad \Delta_n = \int_n^{n+1} \frac{c_0\, dt}{(t + e)\log(t + e)[\log\log(t + e)]^2}.$$

For $N \geq 1$, define the exact inactive gain-box support bound

$$M_N(\delta) := \max_{n\leq N} \sup_{a\in\mathcal{A}} \left\{\ell_a(q_n, q_n^2) + \delta(2q_n/a + q_n^2/a^2)\right\}.$$

This supremum is a finite calculation. For each $q = q_n$, put $f_q(a) = 2q(1 + \delta)/a - q^2(1 - \delta)/a^2$ and $a^* = q(1 - \delta)/(1 + \delta)$. If $a^* \leq 1$, its maximum is $f_q(1)$. Otherwise, set $j = \lfloor \exp(\exp(c_0/(a^* - 1))) - e\rfloor \geq 1$; the maximum is $\max\{f_q(a_j), f_q(a_{j+1})\}$, since $f_q$ increases up to $a^*$ and decreases thereafter.

Because $\ell_a(q_n, q_n^2) \geq 0$, this is the supremum of the absolute functional over all independent gains. The finite sufficient condition is

$$M_N(\delta) < 1 - 3\delta. \tag{52}$$

Choose any $\gamma$ in $((1 - 3\delta)^{-1}, M_N(\delta)^{-1})$. At $q = a_n$, the functional $\ell_{a_n}$ of the perturbed, scaled response is at least $\gamma(1 - 3\delta) > 1$; at every selected midpoint its norm is at most $\gamma M_N(\delta) < 1$. The ordered points $a_{N+1}, q_N, a_N, \ldots, q_1, a_1$, mapped through log, give $2N$ transitions with minimum slack $\mu_N = \min\{\gamma(1 - 3\delta) - 1, 1 - \gamma M_N(\delta)\} > 0$.

For a finite replacement, take three tokens and choose $C > 0$ satisfying

$$\gamma(1 + \delta)\left[\frac{16}{C + 4} + \frac{32}{C + 8}\right] < \mu_N.$$

Give source and query score zero, background score $\log C - hz$ in head $h = 1, 2$, and readouts $\gamma C(1, 0)$ and $\gamma C(0, 1)$. The exact response is $y_C(z) = \gamma(C/(2 + Ce^{-z}), C/(2 + Ce^{-2z}))$. For $q \in [1, 2]$, its coordinate errors before scaling are bounded by $8/(C + 4)$ and $32/(C + 8)$; the preceding inequality bounds the induced norm error under every gain by $\mu_N$. Thus all strict labels survive. The source-relative slopes have magnitude at most two, and Appendix A supplies the attention matrices.

Finally, the exact minimum midpoint deficit is $g_N^{\min} = \min_{n \le N} \Delta_n^2/(4a_n^2)$. The integral formula and $a_n \le 2$ give the finite bound

$$g_N^{\min} \ge \frac{c_0^2}{16(N+1+e)^2[\log(N+1+e)]^2[\log\log(N+1+e)]^4}.$$

Since $M_N(\delta) \le 1 - \min_{n \le N} g_n + 8\delta$, $\min_{n \le N} g_n > 11\delta$ suffices for (52). For sufficiently small fixed $c > 0$, take $N = \lfloor c\delta^{-1/2}/[D(\log D)^2] \rfloor$ as $D \to \infty$. Then $\log N \sim D/2$, $\log\log N \sim \log D$, and decreasing $c$ makes this sufficient condition hold. Hence $2N = \Omega(\delta^{-1/2}/[D(\log D)^2])$. The norm, three-token context, and score slopes remain fixed in this limit. □