GEOMETRIC CAPACITY OF TOP-k AND SLIDING-WINDOW ATTENTION: A THEORETICAL ANALYSIS VIA TROPICAL GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Top-k attention and sliding-window attention are two widely used sparse attention mechanisms for reducing the computation and memory cost of long-context models. Although both mechanisms restrict the number of key–value pairs processed by each query, they generate different sets of routing outputs. In this paper, we present the first unified tropical-geometric comparison of their routing capacities under the same post-routing KV activation-memory budget. With fixed keys, we show that both mechanisms partition query space into polyhedral regions in which the selected KV indices remain unchanged. These partitions admit power-diagram representations: global Top-k produces a power diagram indexed by k-element support sets, whereas sliding-window hardmax produces position-dependent local power diagrams indexed by accessible keys. Given N candidate keys and under the stated dimension conditions, the maximum numbers of single-query routing regions are $\Theta(\binom{N}{k})$ for global Top-k and $\Theta(w)$ for a sliding window of width w. For L independently varying queries with full windows, the corresponding sequence-level capacities are $\Theta(\binom{N}{k}^L)$ and $\Theta(w^L)$. When the two mechanisms have the same budget $k = w = B$, their capacity ratio is $\binom{N}{B}/B$ per query and $[\binom{N}{B}/B]^L$ at the sequence level. The proposed geometric framework provides a theoretical basis for comparing and designing sparse attention mechanisms for memory-efficient long-context models.

1 INTRODUCTION

Sparse attention is widely used to improve long-context efficiency in large language models. Sliding-window attention, adopted by Mistral 7B, Gemma 2, and Gemma 3, restricts each query to a local window and reduces attention computation and KV access accordingly (Jiang et al., 2023; Team et al., 2024; 2025). In contrast, DeepSeek-V3.2 uses a lightweight indexer to select globally relevant key–value pairs through content-dependent Top-k routing (DeepSeek-AI et al., 2025). Sliding-window attention and global Top-k attention represent two typical sparse patterns: structured local selection and content-adaptive global selection. Therefore, a systematic comparison of their information-selection abilities under the same resource constraints can help us understand the inherent expressivity of different sparsification strategies and provide a theoretical basis for the design of sparse attention architectures.

Existing studies analyze sparse attention mainly through efficiency, empirical performance, and network-level computational properties. For global Top-k attention, prior work has examined exact sparse selection and learned indexing for reducing memory, computation, and KV-access costs (Gupta et al., 2021; DeepSeek-AI et al., 2025), while theoretical analysis has bounded the distribution and output errors caused by Top-k truncation (Tzachristas et al., 2026). For sliding-window and related hybrid sparse patterns, existing work has studied task performance and computational efficiency (Beltagy et al., 2020), as well as universal approximation, Turing completeness, connectivity conditions, and formal-language computation (Zaheer et al., 2020; Yun et al., 2020; Li & Cotterell, 2026). These results characterize computational cost, approximation quality, and network-level computation, but not the routing partitions induced in query space. Comparing content-adaptive

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.

global selection with fixed local candidate sets therefore requires analyzing the maximum number of full-dimensional routing regions that each mechanism can induce.

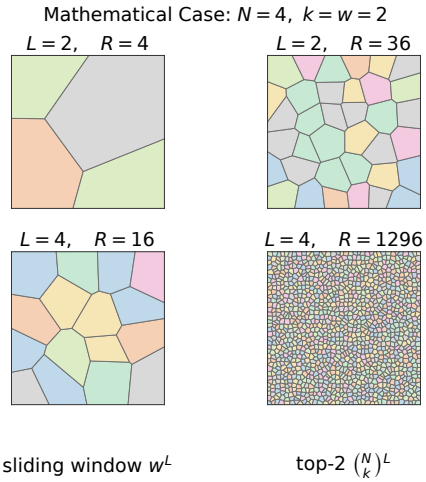


Figure 1: Geometric capacity comparison between sliding-window and Top-2 routing under $N = 4$ and $k = w = 2$ with varying query sequence lengths. Each colored region corresponds to a strict routing region, where the same key groups are selected. The left two panels illustrate sliding-window routing, while the right two panels illustrate Top-2 routing for $L = 2$ and $L = 4$.

For piecewise-linear neural networks, many studies use the maximum number of full-dimensional linear regions in the input space to measure the complexity of the geometric partition induced by a network (Pascanu et al., 2013; Montúfar et al., 2014; Raghu et al., 2017). Related work has established upper and lower bounds, exact counting methods, and growth laws with respect to network depth, width, and input dimension for networks such as ReLU and Maxout networks (Serra et al., 2018; Hanin & Rolnick, 2019; Xiong et al., 2024). Following this idea, we call the maximum number of full-dimensional routing regions that a sparse attention mechanism can produce its routing geometric capacity. A unified theoretical description is still missing for the routing geometric capacities of global Top-k and sliding-window attention and for whether they differ under the same resource constraints.

Unlike ReLU networks, whose linear regions arise from activation patterns, sparse-attention routing is determined by comparisons among affine scores: as the query varies, the maximizing scores and routing label change (Yuan et al., 2025; Vasylenko et al., 2026). Tropical geometry naturally describes these max-affine regions and their tie boundaries (Charisopoulos & Maragos, 2018; Zhang et al., 2018; Alfarra et al., 2022). Applying this framework to the two mechanisms requires handling different combinatorial structures: global Top-k routing is indexed by subsets of keys, sliding-window hardmax has position-dependent candidate sets, and sequence-level analysis must combine the partitions induced by multiple queries. This leads to our central question: ***under fixed keys and the same candidate-KV aggregation budget, what is the maximum number of full-dimensional routing regions that each mechanism can induce, and how do their single-query and sequence-level counts compare?***

To answer this question, we present the first tropical-geometric analysis of both global Top-k attention and sliding-window hardmax. Under fixed keys, we measure routing capacity by the maximum number of full-dimensional linear regions, characterize it at the single-query and sequence levels, and compare the two mechanisms under the same candidate-KV aggregation budget. This analysis clarifies the geometric effects of content-adaptive and fixed local selection and provides a theoretical basis for sparse attention design. Our main contributions are as follows:

- We establish a unified tropical-geometric representation of the two routing mechanisms. Global Top-k partitions query space according to the selected k-key support set, producing a power diagram indexed by combinatorial supports; sliding-window hardmax partitions it according to the winning key in each local window, producing a family of position-dependent local power diagrams.

- We give the maximum number of full-dimensional regions for sliding-window hardmax. When all local cells are nonempty, the single-query region count is $\Theta(w)$. For L independent queries with full windows, the joint region count is $\Theta(w^L)$; for causal windows, it is determined jointly by the window size at each position.

- We give the upper bound $\binom{N}{k}$ for the number of global Top-k routing regions and prove that this bound can be attained under suitable dimension and parameter conditions. For

fixed k, the maximum single-query region count is $\Theta(\binom{N}{k})$, and the joint region count for L independent queries is $\Theta(\binom{N}{k}^L)$.

- When both candidate KV aggregation budgets are B, we compute the ratio between the number of Top-B support-set regions and the number of sliding-window hardmax regions with width B: it is $\binom{N}{B}/B$ for one query and $[\binom{N}{B}/B]^L$ for L independent queries. Figure 1 illustrates how this gap compounds with sequence length when $N = 4$ and $B = k = w = 2$.

2 PRELIMINARIES

In this section, we introduce the fixed-key score model, strict routing regions, geometric capacity, and the two routing mechanisms. Background on tropical geometry, Newton polytopes, and power diagrams is provided in Appendix A.

2.1 MODEL SETTING

Let N denote the number of keys, let d_k denote the dimension of the key and query vectors, and define $[N] := \{1, \dots, N\}$. Let $k_1, \dots, k_N \in \mathbb{R}^{d_k}$ be the key vectors and let $q \in \mathbb{R}^{d_k}$ be a query vector. For each $i \in [N]$, let $\beta_i \in \mathbb{R}$ be a fixed bias and define the corresponding score function as $s_i(q) := \langle q, k_i \rangle + \beta_i$, where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. Throughout this paper, we count regions in the *query space with fixed keys*: the keys and biases are treated as fixed parameters, while the query vector q is allowed to vary.

Let L denote the query-sequence length and write $Q = (q_1, \dots, q_L) \in (\mathbb{R}^{d_k})^L$ for a sequence of L query vectors. The query coordinates $q_1, \dots, q_L$ are allowed to vary independently, while all positions share the same fixed keys and biases. Let $k \in [N]$ denote the number of keys selected by global Top-k attention, and let $w \in [N]$ denote the window width of sliding-window attention. For each query position $t \in \{1, \dots, L\}$, let $W_t \subseteq [N]$ denote its position-dependent candidate window.

We use a two-dimensional example with $N = 4$ and $k = w = 2$ to illustrate the concepts and results developed in this paper.

Example 2.1. Let $N = 4$, $d_k = 2$, and all biases be zero. Set $k_1 = (1, 0), \quad k_2 = (0, 1), \quad k_3 = (-1, -1), \quad k_4 = (0, 0)$. For a query $q = (x, y)$, the four scores are $(s_1(q), s_2(q), s_3(q), s_4(q)) = (x, y, -x - y, 0)$.

Remark 2.2. This example is used only to illustrate the geometric objects. The general results continue to hold for arbitrary N, k, and w, and for arbitrary position-dependent windows W_t.

2.2 GEOMETRIC CAPACITY

Definition 2.3 (Strict routing region)**.** Given a nonempty candidate set $\mathcal{C}$ and an integer $1 \le r \le |\mathcal{C}|$, for any support set $I \subseteq \mathcal{C}$ with $|I| = r$, define its strict routing region as $U_{\mathcal{C},I} := \{q \in \mathcal{Q} : s_i(q) > s_j(q) \text{ for all } i \in I,\ j \in \mathcal{C} \setminus I\}$.

The routing label is the selected support set I. Therefore, changing the order of the scores within I, or within its complement $\mathcal{C} \setminus I$, does not change the routing decision and should not produce additional regions. Definition 2.3 removes this repeated counting by requiring only that every selected score be larger than every unselected score. Moreover, each inequality $s_i(q) > s_j(q)$ defines a strict affine half-space. Hence, $U_{\mathcal{C},I}$ is an intersection of finitely many strict affine half-spaces, so every nonempty strict routing region is open, convex, connected, and full-dimensional in $\mathcal{Q}$, and therefore has positive Lebesgue measure.

Remark 2.4. In Example 2.1, the strict Top-2 region for the support set $I = \{1, 2\}$ is $\mathcal{R}_{\{1,2\}} = \{(x, y) : s_1 > s_3, s_1 > s_4, s_2 > s_3, s_2 > s_4\} = \{(x, y) : x > 0,\ y > 0\}$. There is no need to compare s_1 and s_2, because their order within the support set does not change the routing label. Likewise, there is no need to compare s_3 and s_4. For the window $W_t = \{1, 2\}$, the two strict hardmax regions are $\{x > y\}$ and $\{y > x\}$. The boundary $x = y$ belongs to no strict region because the maximum is tied there.

The number of nonempty strict routing regions measures how many distinct routing decisions are geometrically realizable. To compare model classes, we maximize this number over the routing mechanisms in the class.

Definition 2.5 (Geometric capacity)**.** Let $\mathcal{F}$ be a class of routing mechanisms that can induce polyhedral partitions of the query space $\mathbb{R}^{d_k}$. For any $f \in \mathcal{F}$, let $\mathcal{R}(f)$ denote the set of all nonempty strict routing regions induced by f. We call $\mathcal{C}_{\text{geo}}(\mathcal{F}) := \sup_{f\in\mathcal{F}} |\mathcal{R}(f)|$ the geometric capacity of the model class $\mathcal{F}$. Each region corresponds to a unique routing decision.

This definition provides the common quantity used below to analyze and compare global Top-k attention and sliding-window hardmax.

2.3 Global Top-k and Sliding-Window Attention

We now instantiate the routing framework for the two sparse attention mechanisms considered in this paper. Fix $1 \le k \le N-1$. Global Top-k attention selects the k highest-scoring keys from all N candidates. Its routing label is the unordered support set $I_k(q) := \arg\max_{\substack{I\subseteq[N]\\|I|=k}} \sum_{i\in I} s_i(q), \quad |I_k(q)| = k$. Given value vectors v_i and normalized attention weights $\alpha_i(q)$ on the selected support, the corresponding output is $A_{\text{Top-}k}(q) = \sum_{i\in I_k(q)} \alpha_i(q) v_i$. The routing geometry depends on the selected support $I_k(q)$ rather than on the order of its elements.

Sliding-window attention instead restricts the candidate keys according to the query position. For position $t \in [L]$, let $W_t \subseteq [N], \quad n_t := |W_t| \le w$, where W_t is the accessible key set, n_t is its effective size, and w is the maximum window width. Equivalently, this restriction can be represented by the banded mask

$$M_{tj} := \begin{cases} 0, & j \in \mathcal{W}_t, \\ -\infty, & j \notin \mathcal{W}_t. \end{cases}$$

At temperature $\tau > 0$, the resulting attention weight is $\alpha_{tj}^{(\tau)} = \frac{\exp((s_{tj}+M_{tj})/\tau)}{\sum_{\ell\in\mathcal{W}_t}\exp(s_{t\ell}/\tau)}$. As $\tau \to 0^+$, if the maximum score in W_t is unique, the attention weight concentrates on that key. The zero-temperature sliding-window route is therefore $r_t(q_t) := \arg\max_{j\in W_t} s_j(q_t) = \arg\max_{j\in W_t} \{\langle q_t, k_j\rangle + \beta_j\}$.

Global Top-k routing thus maximizes aggregate affine scores over size-k support sets, whereas sliding-window hardmax maximizes individual affine scores over a position-dependent local candidate set. In both cases, the routing label changes only when competing affine scores become equal. Consequently, their routing regions form polyhedral subdivisions of the query space and can be studied within the common tropical-geometric framework introduced below.

3 Geometric Meaning of the Two Attention Mechanisms

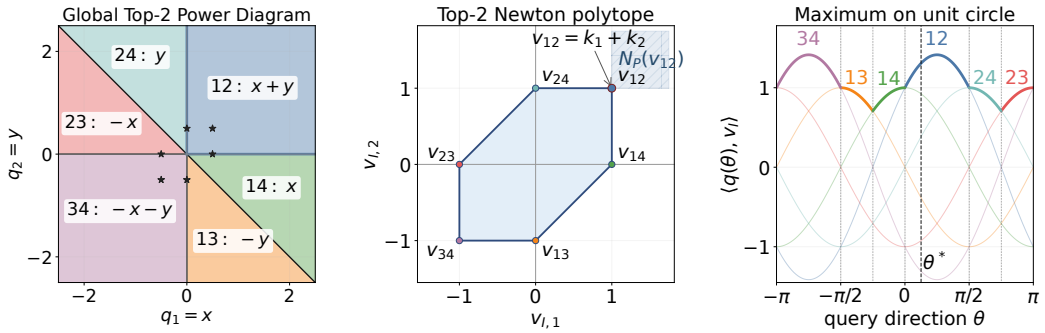
In this section, we interpret global Top-k attention and sliding-window hardmax as power diagrams in query space. This common representation reveals their different routing structures and supports the capacity comparison that follows. Background is provided in Appendix A.

3.1 Top-k Attention: A Global Power Diagram Generated by Support Sets

We first derive its tropical polynomial from the Top-k score function. For each k-element subset $I \subseteq [N]$, define the aggregated slope and aggregated bias by $v_I := \sum_{i\in I} k_i, \quad b_I := \sum_{i\in I} \beta_i$. For any k-element subset $I \subseteq [N]$, we can compute its total score $\sum_{i\in I} s_i(q) = \langle q, v_I\rangle + b_I$. Therefore, the Top-$k$ tropical polynomial can be written as

$$F_k(q) := \max_{\substack{I\subseteq[N]\\|I|=k}} \sum_{i\in I} s_i(q) = \max_{\substack{I\subseteq[N]\\|I|=k}} \big\{\langle q, v_I\rangle + b_I\big\}. \tag{1}$$

Proposition 3.1 (Power-diagram representation of Top-k routing)**.** *For each k-element support set $I \subseteq [N]$, define $p_I := \frac{v_I}{2}, \quad \omega_I := b_I + \frac{\|v_I\|^2}{4}$. Then the strict Top-$k$ routing region of support set I is exactly the open power cell of the weighted site (p_I, ω_I): $\mathcal{R}_I = \big\{q \in \mathbb{R}^{d_k} : \pi_I(q) < \pi_J(q),\ \forall J \ne I,\ |J| = k\big\}$. Therefore, the closures of all strict Top-k routing regions form a power diagram indexed by k-element support sets.*



(a) Global Top-2 power diagram. (b) Top-2 Newton polytope. (c) Top-2 tropical envelope.

Figure 2: (a) Global power diagram induced by Top-2 routing. Each colored cell corresponds to a query region selecting the same key pair, while black boundaries indicate equal total scores between adjacent supports. (b) Newton polytope of Top-2 routing, where each vertex corresponds to a key pair and its value is the sum of the key vectors. (c) Score variation of all key pairs as the query traverses the unit circle. Bold colored arcs denote the key pairs achieving the maximum total score and selected by Top-2 routing.

Proposition 3.1 gives a nearest-site interpretation of Top-k routing. Each k-element support set is represented by one virtual weighted site (p_I, ω_I). Assigning a query to the site with the smallest power distance is exactly equivalent to selecting the support set with the largest total attention score. Thus, one power cell contains all queries that produce the same Top-k support set, and crossing a cell boundary changes at least one selected key. This representation is used for analysis; it does not require an implementation to enumerate or construct all $\binom{N}{k}$ sites. A proof is provided in Appendix B.2.

Remark 3.2. If $b_I = c - \|v_I\|_2^2/4$ for every k-element support set I, then all power weights satisfy $\omega_I = c$. In this case, comparing power distances reduces to comparing ordinary Euclidean distances, so Top-k routing is represented by an ordinary Voronoi diagram with sites $p_I = v_I/2$.

Example 3.3. Figure 2a visualizes the routing behavior of the running Top-2 example. Because there are $\binom{4}{2} = 6$ possible two-key support sets, the diagram contains six virtual sites and six routing regions. Each colored region contains the queries that select the same pair of keys, and the black lines mark queries at which two support sets have equal total scores. For example, queries in the first quadrant select $I = \{1, 2\}$ because the corresponding aggregate score $x + y$ is maximal. Crossing a black boundary changes the selected pair.

The power diagram describes which support set is selected for each query, but it does not yet explain the combinatorial structure of all support sets. A k-element support set can also be encoded by a k-hot indicator vector $\mathbf{1}_I \in \{0, 1\}^N$. The convex hull of all such indicator vectors is the hypersimplex $\Delta_{k,N}$, and the map $K^\top$ sends $\mathbf{1}_I$ to the aggregated slope $K^\top \mathbf{1}_I = \sum_{i \in I} k_i = v_I$. This gives a compact geometric description of all Top-k candidates and motivates the next proposition.

Proposition 3.4 (Hypersimplex representation of the Top-k Newton polytope)**.** *The Newton polytope of the Top-k tropical polynomial is* $P_{N,k} := \operatorname{conv}\left\{\sum_{i \in I} k_i : I \subseteq [N],\ |I| = k\right\}$. *Let* $\Delta_{k,N} := \operatorname{conv}\{\mathbf{1}_I : I \subseteq [N], |I| = k\} \subseteq \mathbb{R}^N$ *be the (k, N)-hypersimplex, and let the linear map* $K^\top : \mathbb{R}^N \to \mathbb{R}^{d_k}$ *satisfy* $K^\top e_i = k_i$. *Then* $P_{N,k} = K^\top \Delta_{k,N}$.

Proposition 3.4 shows that the Top-k Newton polytope is obtained by projecting the hypersimplex through the key matrix. In computational terms, the hypersimplex contains one vertex for every possible k-hot selection pattern, while $K^\top$ converts that pattern into the sum of its selected keys. The projected vertices therefore encode the candidate support sets that can define routing regions. When the aggregate biases b_I are constant, the normal cones of the exposed vertices give the routing regions directly; for general biases, the corresponding weighted regions are described by the power diagram in Proposition 3.1. A proof is provided in Appendix B.3.

Example 3.5. For the simple example, the six possible two-key selections produce the six aggregated slopes shown in Figure 2b. Their convex hull is the hexagon $P_{4,2}$. Each vertex represents one

candidate support set; for example, $v_{12} = k_1 + k_2$ represents selecting keys 1 and 2. Its normal cone contains the query directions for which this support is optimal.

Figure 2c shows the same routing process from the score viewpoint. As the query moves along the unit circle, the six aggregate scores change, and the upper envelope records the largest one. A change in the active segment of the envelope corresponds exactly to a change in the selected support set.

3.2 Sliding-Window Attention: A Family of Local Power Diagrams Generated by Banded Position Indices

Sliding-window hardmax has a simpler candidate structure. At position t, the mask first removes all keys outside W_t, and hardmax then selects one key from the remaining candidates. Thus, unlike global Top-k routing, whose sites represent entire support sets, each local site represents one accessible key.

Proposition 3.6 (Local power-diagram family representation of sliding-window attention)**.** *For each position $t \in [N]$ and each key $j \in \mathcal{W}_t$ in the window, define $p_{t,j} := \frac{k_j}{2}, \quad \omega_{t,j} := \beta_{tj} + \frac{\|k_j\|^2}{4}$. Then the following statements hold:*

1. *The strict hardmax routing region of key j at position t is exactly the open power cell of the weighted site $(p_{t,j}, \omega_{t,j})$:*

$$\mathcal{R}_{t,j} = \left\{ q_t \in \mathbb{R}^d : \pi_{t,j}(q_t) < \pi_{t,\ell}(q_t),\ \forall \ell \in \mathcal{W}_t \setminus \{j\} \right\},$$

   where $\pi_{t,j}(q_t) := \|q_t - p_{t,j}\|^2 - \omega_{t,j}$. Therefore, the closures of all nonempty strict hardmax routing regions at position t form the power diagram generated by the local site set $\mathcal{S}_t := \{(p_{t,j}, \omega_{t,j}) : j \in \mathcal{W}_t\}$

2. *As the position t changes, sliding-window attention generates a family of local power diagrams controlled by the position index,*

$$\mathcal{D}_{\mathrm{SWA}} := \{\mathrm{Pow}(\mathcal{S}_t) : t = 1, \ldots, L\}.$$

   If $q_1, \ldots, q_L$ are treated as independent variables, the joint routing partition of the whole layer is the Cartesian product of these local power diagrams.

Proposition 3.6 states that each query position has its own local routing diagram. At position t, the power cell of key j contains exactly the queries for which j has the largest score among the keys in W_t. Moving to another position may change the window and therefore changes the set of competing sites. Under the product-space model for $(q_1, \ldots, q_L)$, a layer-level routing label is a tuple of local decisions, so the joint partition is the Cartesian product of the local diagrams. A proof is provided in Appendix B.4.

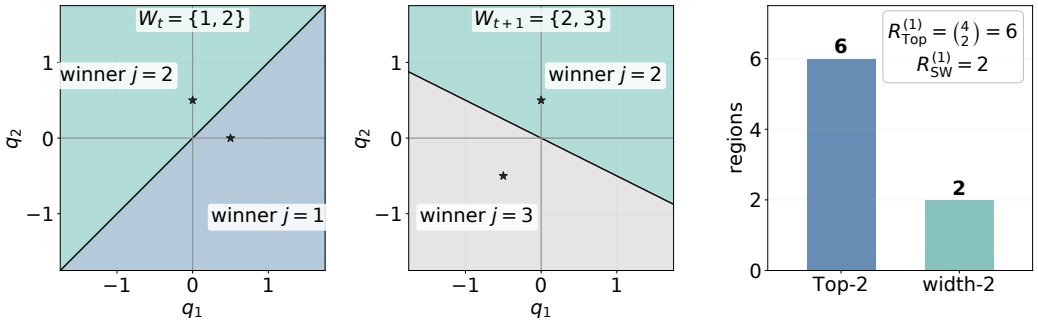
Example 3.7. For the example, the local power diagrams induced by sliding-window hardmax at two neighboring positions are shown in Figure 3a. The candidate window in the left panel is $W_t = 1, 2$, and that in the right panel is $W_{t+1} = 2, 3$. The stars represent the local weighted sites $p_{t,j} = k_j/2$, and the black lines represent the decision boundaries where the two candidate keys in a window have equal scores.

4 Geometric Capacity of the Two Attention Routing Mechanisms

In this section, we analyze the geometric capacities of global Top-k attention and sliding-window hardmax, and compare them under the same post-routing KV activation-memory budget.

4.1 Upper Bound on Strict Routing Regions of Top-k Attention

For global Top-k attention, a routing label is a k-element subset $I \subseteq [N]$. It is selected when every key in I scores higher than every key outside I. Hence, $\mathcal{R}_I = \bigcap_{\substack{i \in I \\ j \in [N] \setminus I}} \{q \in \mathbb{R}^{d_k} : \langle q, k_i - k_j \rangle + \beta_i - \beta_j > 0\}$. Because $\mathcal{R}_I$ is convex, one support set can produce at most one strict routing region. The number of possible support labels therefore immediately gives an upper bound.



(a) Sliding window power diagram.

(b) Capacity comparison.

Figure 3: (a) Local power diagrams of sliding-window attention with $N = 4$, $w = 2$. In the left panel, the window contains the first and second keys; in the right panel, it contains the second and third keys. Colored regions indicate query regions where a particular key attains the highest score, while the black line marks queries where the two keys have equal scores.(b) Bar chart comparing the numbers of strict routing regions under the single-query setting for Top-2 attention and sliding-window attention with window size 2.

Theorem 4.1 (Upper bound on Top-k routing regions)**.** *The number of single-query strict Top-k routing regions induced by any fixed keys satisfies $R^{(1)}_{\text{Top-}k} \le \binom{N}{k}$. In the fixed-key, independent-query model, the number of joint regions for the whole sequence satisfies $R_{\text{Top-}k,\text{seq}} \le \left(R^{(1)}_{\text{Top-}k}\right)^L \le \binom{N}{k}^L$.*

Theorem 4.1 can be understood directly from the router output. Each query returns an unordered set of k key indices, giving $\binom{N}{k}$ possible routing labels. For fixed keys, all queries producing the same label form a single convex region, so the number of regions cannot exceed the number of labels. For L independent queries, the sequence-level output is an L-tuple of support sets, yielding at most $\binom{N}{k}^L$ joint regions. The proof is provided in Appendix B.5.

4.2 Constructive Lower Bound for Top-k Attention

The upper bound counts all possible support labels, but it does not show that every label can actually occur. The next theorem gives one fixed key configuration for which each k-element subset is selected on a nonempty open set of queries.

Theorem 4.2 (Explicit constructive lower bound for Top-k routing)**.** *If $d_k \ge 2\min\{k, N-k\}$, where d_k is the dimension of the query and key spaces. Then there is a fixed key and query configuration for which every k-element subset $I \subseteq [N]$ is the unique Top-k support set in some region. Therefore, $R^{(1)}_{\text{Top-}k} = \binom{N}{k}$. When the sequence queries are independent for the same set of keys, $R_{\text{Top-}k,\text{seq}} = \binom{N}{k}^L$.*

Theorem 4.2 shows that this upper bound is attainable. When $d_k \ge 2\min\{k, N-k\}$, one fixed key configuration allows every k-element subset to be selected for a nonempty range of queries. Thus, all $\binom{N}{k}$ routing labels are reachable, and the sequence-level capacity is $\binom{N}{k}^L$ for L independent queries. The construction uses the moment curve and squared polynomials to make any target support set score strictly higher than its complement. The proof is provided in Appendix B.6.

4.3 Upper Bound on Strict Routing Regions of Sliding-Window Hardmax

We next analyze the number of strict routing regions of sliding-window hardmax attention.

Theorem 4.3 (Upper bound on sliding-window hardmax routing regions)**.** *The number of single-query strict routing regions at position t satisfies $R^{(1)}_{\text{SWA},t} \le n_t \le w$. In the fixed-key, independent-*

query model, the number of joint regions for the whole sequence satisfies $R^{(1)}_{\mathrm{SWA,seq}} \le \prod_{t=1}^{L} n_t \le w^L$.

Theorem 4.3 follows from the output space of the local router. At position t, the router returns one key index from the n_t candidates in $\mathcal{W}_t$, so it can produce at most n_t routing regions. For L independent queries, a sequence-level output is a tuple of L local key indices, giving at most $\prod_{t=1}^{L} n_t$ joint regions. When every window has width w, this upper bound becomes w^L. The proof is provided in Appendix B.7.

4.4 Constructive Lower Bound for Sliding-Window Hardmax

Theorem 4.3 bounds the capacity by the number of candidate keys in each window, but it remains to show that every candidate can be selected on a nonempty query region. We now construct a shared set of fixed keys that attains this upper bound, thereby establishing the exact geometric capacity of sliding-window hardmax.

Theorem 4.4 (Explicit constructive lower bound for sliding-window hardmax routing)**.** *If $d_k \ge 2$, then there is a set of shared fixed keys such that all n_t candidate keys at every position t correspond to nonempty strict routing regions. Therefore, $R^{(1)}_{\mathrm{SWA},t} = n_t, \quad R^{(1)}_{\mathrm{SWA,seq}} = \prod_{t=1}^{L} n_t$. If the window width at every position is w, then $R^{(1)}_{\mathrm{SWA},t} = w, \quad R^{(1)}_{\mathrm{SWA,seq}} = w^L$.*

Theorem 4.4 shows that every local routing output can be realized when $d_k \ge 2$. The construction arranges the affine score functions so that each candidate key becomes the unique winner over a nonempty range of queries. Consequently, all n_t local labels are reachable at position t, making the upper bounds n_t and $\prod_{t=1}^{L} n_t$ tight. The construction uses tangent lines to the parabola $f(\tau) = \tau^2/4$, with each line appearing on the upper envelope over a distinct interval. The proof is provided in Appendix B.8.

In summary, under the stated dimension conditions, the maximum single-query geometric capacities of global Top-k attention and sliding-window hardmax are $R^{(1),\max}_{\text{Top-}k} = \Theta(\binom{N}{k}), \quad R^{(1),\max}_{\mathrm{SWA}} = \Theta(w)$. For L independent queries with full windows, their maximum sequence-level capacities are $\Theta(\binom{N}{k}^L)$ and $\Theta(w^L)$, respectively.

4.5 Geometric Capacity Comparison of the Two Attention Mechanisms

Having derived the geometric capacities of both mechanisms, we now compare them under a common resource constraint. We first define the post-routing KV activation-memory budget and then evaluate their single-query and sequence-level capacity ratios under the matched budget.

4.5.1 Post-Routing KV Activation-Memory Budget

We measure the resource constraint by the number of KV pairs passed to the attention scoring and aggregation stage.

Definition 4.5 (Post-routing KV activation-memory budget)**.** Let $\widetilde{S}^{\rho}_t(q_t)$ denote the KV support set sent by routing mechanism ρ to the attention scoring and aggregation stage at position t. If both the key and value dimensions are d_k, and each scalar occupies c bytes, define the post-routing KV activation-memory budget at position t as $\mathcal{M}^{\rho}_t(q_t) := 2cd_k \left|\widetilde{S}^{\rho}_t(q_t)\right|$. For a query sequence of length L, define the total budget as $\mathcal{M}^{\rho}(q_{1:L}) := 2cd_k \sum_{t=1}^{L} \left|\widetilde{S}^{\rho}_t(q_t)\right|$. This budget does not include the retrieval, sorting, and indexing costs required to construct $\widetilde{S}^{\rho}_t(q_t)$ from the full candidate set.

Global Top-k routing passes k selected KV pairs to the attention stage, whereas sliding-window attention accesses the w KV pairs in its local window. Setting $k = w = B$ therefore gives both mechanisms the same per-query post-routing KV activation-memory budget. Sliding-window hardmax still accesses all B local candidates before selecting the winner, while the search, sorting, and indexing costs used to retrieve the global Top-B support are outside this budget. Having matched the number of KV pairs processed per query, we can now compare how the two routing structures use this budget.

4.5.2 Comparison of the Two Attention Mechanisms

Under the matched budget $k = w = B$, sliding-window hardmax selects one winner from B fixed local candidates, whereas global Top-B selects a B-element support set from all N keys. The following theorem quantifies the resulting difference in their single-query and sequence-level geometric capacities.

Theorem 4.6 (Geometric-capacity comparison under the same post-routing KV budget)**.** *Suppose every query is given the same set of N fixed candidate keys and the query and key dimensions satisfy $d_k \geq 2\min\{B, N-B\}$. Then Top-B attention and sliding-window attention with width B have the same per-query support budget, where $1 \leq B < N$. For any fixed position t, $\frac{R^{(1),\max}_{\text{Top-}B}}{R^{(1),\max}_{\text{SWA},t}} = \frac{\binom{N}{B}}{B}$. If $q_1, \ldots, q_L$ are independent query variables, then the joint geometric capacities for the whole sequence satisfy*

$$\frac{R^{\max}_{\text{Top-}B,\text{seq}}}{R^{\max}_{\text{SWA,seq}}} = \left(\frac{\binom{N}{B}}{B}\right)^L$$

Theorem 4.6 compares the geometric capacities of two routing mechanisms under the same budget B. Sliding-window hardmax selects one key from B local candidates and therefore has at most B routing regions, whereas global Top-B selects a support set from N keys and can realize $\binom{N}{B}$ regions. Thus, their single-query capacity ratio is $\binom{N}{B}/B$, and this difference compounds across independent queries. Figure 3b illustrates the single-query case with $N = 4$ and $B = 2$, where global Top-2 realizes six routing regions and sliding-window hardmax realizes two. The proof is provided in Appendix B.9.

Example 4.7. Let $N = 4$ and $k = w = B = 2$. The single-query capacities of global Top-2 and sliding-window hardmax are 6 and 2, with a ratio of 3. As shown in Figure 1, their sequence-level capacities are 36 and 4 for $L = 2$, and 1296 and 16 for $L = 4$, giving ratios of 9 and 81, respectively. Each colored cell represents a distinct routing decision.

5 Conclusion

In this work, we give the first tropical-geometric analysis of global Top-k attention and sliding-window hardmax. Global Top-k induces a power diagram indexed by k-element support sets, whereas sliding-window hardmax induces position-dependent local power diagrams indexed by accessible keys. Under the stated dimension conditions, the maximum single-query capacities of global Top-k and sliding-window hardmax are $\Theta(\binom{N}{k})$ and $\Theta(w)$, while their capacities for L independent queries are $\Theta(\binom{N}{k}^L)$ and $\Theta(w^L)$. Under the same budget $k = w = B$, their capacity ratios are $\binom{N}{B}/B$ and $[\binom{N}{B}/B]^L$. These results quantify the combinatorial advantage of global content-adaptive routing and inform the design of sparse attention for long-context models.

These formulas provide a budget-aware criterion for sparse-attention design and selection. Given the context size N, sequence length L, and KV budget B, they quantify the routing diversity available to different mechanisms before model training. This can guide the use of global Top-k routing for tasks requiring content-dependent long-range retrieval, sliding windows for workloads dominated by local dependencies and predictable memory access, and hybrid architectures that allocate local and global routing to different heads or layers. The results also provide a theoretical basis for selecting k and w and narrowing the architecture search space before empirical tuning.

Our analysis assumes fixed keys and independently varying queries; in full self-attention, queries and keys depend on shared inputs, producing coupled routing decisions and generally quadratic boundaries in the input space. Moreover, the condition $d_k \geq 2\min\{k, N-k\}$ used to attain the Top-k upper bound is sufficient but may not be minimal. Future work may extend the analysis to dynamic keys using semialgebraic geometry, determine tighter dimension requirements, and study multi-head attention through common refinements of multiple routing partitions.

AI USE STATEMENT

In this work, we did not use generative AI tools for tasks requiring disclosure such as developing the theoretical framework, formulating mathematical claims, proving theorems, designing experiments, implementing methods, processing data, or interpreting empirical results. We used generative AI tools only for manuscript writing support, including language polishing, improving readability, condensing text, and reorganizing presentation. All AI-assisted text was reviewed and edited by the authors. The authors verified the mathematical statements, proofs, and all scientific claims independently, and take full responsibility for the final content of this work.

ETHICS STATEMENT

This work is theoretical and methodological, and does not involve human subjects, private or sensitive data, dataset release, surveillance, security-sensitive applications, or deployment in high-stakes domains. We do not identify specific ethical concerns beyond standard requirements of research integrity, reproducibility, and accurate reporting.

REPRODUCIBILITY STATEMENT

This work is entirely theoretical and contains no experiments, datasets, or code. All assumptions, definitions, and mathematical conventions are stated explicitly, and complete proofs of all propositions, theorems, and corollaries are provided in the appendix. The results can therefore be verified directly from the stated formulation and proofs.

REFERENCES

Motasem Alfarra, Adel Bibi, Hasan Hammoud, Mohamed Gaafar, and Bernard Ghanem. On the decision boundaries of neural networks: A tropical geometry perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5027–5037, 2022.

Franz Aurenhammer. Power diagrams: properties, algorithms and applications. *SIAM journal on computing*, 16(1):78–96, 1987.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Pascal Bergsträßer, Chris Köcher, Anthony W Lin, and Georg Zetzsche. The power of hard attention transformers on data sequences: A formal language theoretic perspective. *Advances in Neural Information Processing Systems*, 37:96750–96774, 2024.

Vasileios Charisopoulos and Petros Maragos. A tropical approach to neural networks with piecewise linear activations. *arXiv preprint arXiv:1805.08749*, 2018.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Erhang Li, Fangqi Zhou, Fangyun Lin, Fucong Dai, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoran Wei, Haowei Zhang, Haowen Luo, Haozhe Ji, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, Jialiang Huang, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jingchang Chen, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jinhua Zhu, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Kexin Huang, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Liang Zhao, Liangsheng Yin, Lihua Guo, Lingxiao Luo, Linwang Ma, Litong Wang, Liyue Zhang, M. S. Di, M. Y Xu, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Panpan Huang, Peixin Cong, Peiyi Wang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang,

Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, S. H. Liu, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Songyang Zhou, Tao Ni, Tao Yun, Tian Pei, Tian Ye, Tianyuan Yue, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjun Gao, Wentao Zhang, Xi Gao, Xiangwen Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyuan Li, Xu Chen, Xuecheng Su, Xuehai Pan, Xuheng Lin, Xuwei Fu, Y. Q. Wang, Yang Zhang, Yanhong Xu, Yanru Ma, Yao Li, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yiliang Xiong, Ying He, Ying Zhou, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuduan Wang, Yue Gong, Yuhan Wu, Yuheng Zou, Yukun Li, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehua Zhao, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhiyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Zizheng Pan, Zongqing Yao, Bei Feng, Hui Li, J. L. Cai, Jiaqi Ni, Lei Xu, Meng Li, Ning Tian, R. J. Chen, R. L. Jin, S. S. Li, Shuang Zhou, Tianyu Sun, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xinnan Song, Xinyi Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, Dongjie Ji, Jian Liang, Jianzhong Guo, Jin Chen, Leyi Xia, Miaojun Wang, Mingming Li, Peng Zhang, Ruyi Chen, Shangmian Sun, Shaoqing Wu, Shengfeng Ye, T. Wang, W. L. Xiao, Wei An, Xianzu Wang, Xiaowen Sun, Xiaoxiang Wang, Ying Tang, Yukun Zha, Zekai Zhang, Zhe Ju, Zhen Zhang, and Zihua Qu. Deepseek-v3.2: Pushing the frontier of open large language models, 2025. URL https://arxiv.org/abs/2512.02556.

Aditya Desai, Kumar Agrawal, Shuo Yang, Alejandro Cuadron Lafuente, Luis Gaspar Schroeder, Matei Zaharia, Joseph E Gonzalez, and Ion Stoica. vattention: Verified sparse attention via sampling. In *International Conference on Learning Representations*, volume 2026, pp. 52253–52284, 2026.

Ankit Gupta, Guy Dar, Shaya Goodman, David Ciprut, and Jonathan Berant. Memory-efficient transformers via top-k attention. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pp. 39–52, 2021.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020.

Boris Hanin and David Rolnick. Complexity of linear regions in deep networks. In *International Conference on Machine Learning*, pp. 2596–2604. PMLR, 2019.

Baran Hashemi, Kurt Pasque, Chris Teska, and Ruriko Yoshida. Tropical attention: Neural algorithmic reasoning for combinatorial algorithms. *Advances in Neural Information Processing Systems*, 38:98631–98656, 2026.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Jiaoda Li and Ryan Cotterell. Characterizing the expressivity of local attention in transformers. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37485–37507, 2026.

Diane Maclagan and Bernd Sturmfels. *Introduction to tropical geometry*, volume 161. American Mathematical Society Providence, RI, 2015.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. *Advances in neural information processing systems*, 27, 2014.

Razvan Pascanu, Guido Montufar, and Yoshua Bengio. On the number of response regions of deep feed forward networks with piece-wise linear activations. *arXiv preprint arXiv:1312.6098*, 2013.

Tuan Anh Pham and Vikas Garg. What do graph neural networks learn? insights from tropical geometry. *Advances in Neural Information Processing Systems*, 37:10988–11020, 2024.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *international conference on machine learning*, pp. 2847–2854. PMLR, 2017.

Thiago Serra, Christian Tjandraatmadja, and Srikumar Ramalingam. Bounding and counting linear regions of deep neural networks. In *International conference on machine learning*, pp. 4558–4566. PMLR, 2018.

Georgios Smyrnis and Petros Maragos. Multiclass neural network minimization via tropical newton polytope approximation. In *International Conference on Machine Learning*, pp. 9068–9077. PMLR, 2020.

Ye Su and Yong Liu. Geometric capacity of transformers: A tropical geometry perspective, 2026. URL https://arxiv.org/abs/2604.14727.

Ye Su, Huayi Tang, Zixuan Gong, and Yong Liu. Sparsity is combinatorial depth: Quantifying moe expressivity via tropical geometry. *arXiv preprint arXiv:2602.03204*, 2026.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas

Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL `https://arxiv.org/abs/2503.19786`.

Georgios Tzachristas, Lei Deng, Ioannis Tzachristas, Gong Zhang, and Renhai Chen. A mathematical theory of top-k sparse attention via total variation distance. In *2026 IEEE International Symposium on Information Theory (ISIT)*, pp. 1–6. IEEE, 2026.

Pavlo Vasylenko, Hugo Pitorro, Andre Martins, and Marcos V Treviso. Long-context generalization with sparse attention. In *International Conference on Learning Representations*, volume 2026, pp. 15256–15303, 2026.

Huan Xiong, Lei Huang, Wenston JT Zang, Xiantong Zhen, Guo-Sen Xie, Bin Gu, and Le Song. On the number of linear regions of convolutional neural networks with piecewise linear activations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):5131–5148, 2024.

Di Xiu, Hongyin Tang, Bolin Rong, Lizhi Yan, Jingang Wang, Yifan Lu, and Xunliang Cai. A preliminary study on the promises and challenges of native top-k sparse attention. *arXiv preprint arXiv:2512.03494*, 2025.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Yuxing Wei, Lean Wang, Zhiping Xiao, et al. Native sparse attention: Hardware-aligned and natively trainable sparse attention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23078–23097, 2025.

Chulhee Yun, Yin-Wen Chang, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. O (n) connections are expressive enough: Universal approximability of sparse transformers. *Advances in Neural Information Processing Systems*, 33:13783–13794, 2020.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

Liwen Zhang, Gregory Naitzat, and Lek-Heng Lim. Tropical geometry of deep neural networks. In *International Conference on Machine Learning*, pp. 5824–5832. PMLR, 2018.

A Background on Tropical Geometry and Power Diagrams

This appendix collects the geometric background used in the main analysis. Both global Top-k routing and sliding-window hardmax can be represented by maximization over finitely many affine functions. Tropical polynomials describe these max-affine functions and their boundaries, Newton polytopes encode their combinatorial structure, and power diagrams provide an equivalent Euclidean representation of their polyhedral regions.

Definition A.1 (Tropical semiring (Maclagan & Sturmfels, 2015))**.** The tropical semiring is defined on the set $\mathbb{T} = \mathbb{R} \cup \{-\infty\}$ with the following two operations: $a \oplus b := \max(a, b)$, $a \otimes b := a + b$. Here, $\oplus$ is called tropical addition and $\otimes$ is called tropical multiplication. In this algebraic structure, ordinary addition is replaced by the maximum operation, and ordinary multiplication is replaced by ordinary addition. It is therefore also called max-plus algebra.

Definition A.2 (Tropical polynomial and tropical hypersurface)**.** Let $\mathcal{A} \subset \mathbb{R}^d$ be a finite set of slopes and let $b_a \in \mathbb{R}$ be the coefficient associated with $a \in \mathcal{A}$. A finite generalized tropical polynomial is the max-affine function (Charisopoulos & Maragos, 2018; Zhang et al., 2018)

$$f(x) := \max_{a \in \mathcal{A}} \{\langle a, x\rangle + b_a\}, \quad x \in \mathbb{R}^d. \tag{2}$$

Its tropical hypersurface is the set on which at least two affine terms attain the maximum (Maclagan & Sturmfels, 2015):

$$\mathrm{Trop}(f) := \left\{ x \in \mathbb{R}^d : \left| \arg\max_{a \in \mathcal{A}} \{\langle a, x\rangle + b_a\} \right| \geq 2 \right\}.$$

The function f is convex and piecewise affine. Each nonempty set on which one term uniquely attains the maximum is a full-dimensional linear region, while $\mathrm{Trop}(f)$ contains the boundaries separating these regions. Thus, the strict routing regions defined above are precisely the unique-maximizer regions of an appropriate tropical polynomial.

Definition A.3 (Newton polytope and normal cone)**.** The Newton polytope of f is the convex hull of its slopes (Maclagan & Sturmfels, 2015): $\mathrm{Newt}(f) := \mathrm{conv}(\mathcal{A})$. For a polytope $P \subseteq \mathbb{R}^d$, its support function and the normal cone at a vertex a are

$$h_P(x) := \max_{p \in P} \langle p, x\rangle, \quad N_P(a) := \{x \in \mathbb{R}^d : \langle a, x\rangle \geq \langle p, x\rangle \text{ for all } p \in P\}.$$

If $b_a = 0$ for every $a \in \mathcal{A}$, then $f = h_{\mathrm{Newt}(f)}$, and its full-dimensional linear regions are the interiors of the normal cones of the exposed vertices of $\mathrm{Newt}(f)$. For nonzero coefficients, the lifted points (a, b_a) induce a regular subdivision of the Newton polytope (Maclagan & Sturmfels, 2015). The corresponding translated max-affine regions admit an equivalent representation as a power diagram.

The correspondence between these tropical-geometric objects and single-key hardmax routing is summarized in Table 1. For a candidate set $\mathcal{C} \subseteq [N]$, the query q plays the role of the input variable, while each key k_i and bias β_i determine one affine term. A unique maximizing term gives a strict routing region, whereas a tie between maximal terms lies on the tropical hypersurface.

Table 1 describes single-key hardmax routing. For sliding-window hardmax, all maximizations and comparisons are restricted to $i, j \in W_t$. For global Top-k routing, the affine terms are instead indexed by support sets I, with aggregated slope $a_I = \sum_{i \in I} k_i$ and aggregated intercept $b_I = \sum_{i \in I} \beta_i$. The normal-cone correspondence in the last row applies when all score biases are zero. With nonzero biases, the routing regions are generally translated polyhedra rather than cones and are more naturally represented as power cells, which we introduce next.

Table 1: Correspondence between tropical geometry and single-key hardmax attention routing.

Tropical Geometry	Attention Routing	Mathematical Correspondence
Input variable x	Query vector q	$x = q$
Slope a_i	Key vector k_i	$a_i = k_i$
Intercept b_i	Score bias β_i	$b_i = \beta_i$
Affine term	Attention score	$\langle a_i, x\rangle + b_i = \langle q, k_i\rangle + \beta_i = s_i(q)$
Unique maximal term	Selected key under hardmax	$\arg\max_{i\in\mathcal{C}} s_i(q)$
Full-dimensional linear region	Strict routing region	$\{q : s_i(q) > s_j(q),\ \forall j \in \mathcal{C} \setminus \{i\}\}$
Tropical hypersurface	Routing decision boundary	$\{q : \exists i \neq j,\ s_i(q) = s_j(q) = \max_{\ell\in\mathcal{C}} s_\ell(q)\}$
Newton polytope	Convex hull of candidate keys	$\mathrm{Newt}(f) = \mathrm{conv}\{k_i : i \in \mathcal{C}\}$
Normal cone of a vertex	Closed selection region when $\beta_i = 0$	$\{q : \langle k_i, q\rangle \geq \langle k_j, q\rangle,\ \forall j \in \mathcal{C}\}$

Definition A.4 (Power distance, power cell, and power diagram (Aurenhammer, 1987))**.** Let Λ be a finite index set, and associate each $\lambda \in \Lambda$ with a site $p_\lambda \in \mathbb{R}^d$ and a weight $\omega_\lambda \in \mathbb{R}$. The power distance from $x \in \mathbb{R}^d$ to $(p_\lambda, \omega_\lambda)$ is $\pi_\lambda(x) := \|x - p_\lambda\|_2^2 - \omega_\lambda$. The power cell of $(p_\lambda, \omega_\lambda)$ is

$$\mathrm{Pow}_\lambda := \{x \in \mathbb{R}^d : \pi_\lambda(x) \leq \pi_\mu(x) \text{ for all } \mu \in \Lambda\}.$$

The collection of all nonempty power cells is the power diagram, also called a weighted Voronoi diagram: $\mathrm{Pow}\big(\{(p_\lambda, \omega_\lambda)\}_{\lambda\in\Lambda}\big) := \{\mathrm{Pow}_\lambda : \lambda \in \Lambda,\ \mathrm{Pow}_\lambda \neq \varnothing\}$.

The connection to tropical geometry follows from a standard transformation between max-affine functions and power diagrams (Aurenhammer, 1987).

Lemma A.5 (Equivalence of max-affine partitions and power diagrams)**.** *For the max-affine function f in Eq.* (2)*, associate each affine term with the weighted site*

$$(p_a, \omega_a) := \left(\frac{a}{2},\ b_a + \frac{\|a\|_2^2}{4}\right), \quad a \in \mathcal{A}. \tag{3}$$

Then the closed maximality region of the a-th term,

$$C_a := \{x \in \mathbb{R}^d : \langle a, x\rangle + b_a \geq \langle c, x\rangle + b_c \textit{ for all } c \in \mathcal{A}\},$$

is exactly the power cell Pow_a *of (p_a, ω_a). If C_a is full-dimensional, its interior is the strict region on which the a-th term is the unique maximizer.*

See Appendix B.1 for the proof. This equivalence also underlies the power-diagram characterization of fixed-key, zero-temperature self-attention (Su & Liu, 2026).

Remark A.6. For Example 2.1, the tropical polynomial associated with global Top-2 routing is $F_2(x, y) := \max\{x + y, -y, x, -x, y, -x - y\}$. The six terms correspond, in order, to the support sets $12, 13, 14, 23, 24, 34$. For a support set I, let $a_I := \sum_{i\in I} k_i$ denote its aggregated slope. The six slopes in this example are $(a_{12}, a_{13}, a_{14}, a_{23}, a_{24}, a_{34}) = \big((1, 1), (0, -1), (1, 0), (-1, 0), (0, 1), (-1, -1)\big)$. Their convex hull is the hexagon $P_{4,2} := \mathrm{conv}\{a_I : I \subseteq [4],\ |I| = 2\}$. Every a_I is an exposed vertex of $P_{4,2}$, and the interior of its normal cone is the corresponding strict Top-2 region. Equivalently, Lemma A.5 identifies the closure of each region with a power cell. The tropical hypersurface is the union of the lines $x = 0$, $y = 0$, and $x + y = 0$, which divide the query plane into the six open cones shown in Figure 2(b).

In summary, affine terms encode routing labels, unique-maximizer regions encode strict routing regions, tropical hypersurfaces record routing boundaries, and power cells provide the corresponding weighted Voronoi representation. These facts provide the geometric background for the unified routing analysis in the main text.

B Proofs of Lemma, Theorems and Propositions

This section gives the complete proofs of the lemmas, propositions, and theorems in the main text.

B.1 Lemma A.5 (Equivalence Between Max-Affine Partitions and Power Diagrams)

Proof. By Eq. (3),

$$\pi_a(x) = \left\| x - \frac{a}{2} \right\|^2 - b_a - \frac{\|a\|^2}{4} = \|x\|^2 - \langle a, x \rangle - b_a.$$

The common term $\|x\|^2$ is independent of the index a. Therefore, minimizing $\pi_a(x)$ is equivalent to maximizing $\langle a, x \rangle + b_a$. □

B.2 Proposition 3.1 (Power-Diagram Representation of Top-k Routing)

Proof. We first prove that I is the unique maximal term in the tropical polynomial of the attention mechanism if and only if I is the strict Top-k active set for query q.

First, suppose that $q \in \mathcal{R}_I$, that is,

$$s_i(q) > s_j(q), \quad \forall i \in I,\ j \notin I.$$

Take any other support set J with $|J| = k$ and $J \neq I$. Clearly,

$$|I \setminus J| = |J \setminus I| \geq 1.$$

Moreover, for any $i \in I \setminus J$ and $j \in J \setminus I$, we have $s_i(q) > s_j(q)$. Therefore,

$$\sum_{i \in I \setminus J} s_i(q) > \sum_{j \in J \setminus I} s_j(q).$$

Adding the common part $I \cap J$ gives

$$\sum_{i \in I} s_i(q) > \sum_{j \in J} s_j(q).$$

Because this inequality holds for every $J \neq I$, I is the unique maximal term in Eq. (1).

Conversely, suppose that I is the unique maximal term in Eq. (1). If I is not the strict Top-k active set, then there exist $i \in I$ and $j \notin I$ such that

$$s_i(q) \leq s_j(q).$$

Let

$$J := (I \setminus \{i\}) \cup \{j\}.$$

Clearly, $|J| = k$ and $J \neq I$, and

$$\sum_{\ell \in J} s_\ell(q) = \sum_{\ell \in I} s_\ell(q) - s_i(q) + s_j(q) \geq \sum_{\ell \in I} s_\ell(q).$$

This contradicts the assumption. Hence,

$$s_i(q) > s_j(q), \quad \forall i \in I,\ j \notin I,$$

so $q \in \mathcal{R}_I$. In summary,

$$\mathcal{R}_I = \left\{ q \in \mathbb{R}^{d_k} : \langle q, v_I \rangle + b_I > \langle q, v_J \rangle + b_J,\ \forall J \neq I,\ |J| = k \right\}. \tag{4}$$

Finally, we prove that the closures of all nonempty strict Top-k routing regions form a power diagram indexed by k-element support sets.

Eq. (1) is a max-affine function with slope v_I and bias b_I. By Lemma A.5, when

$$p_I = \frac{v_I}{2}, \quad \omega_I = b_I + \frac{\|v_I\|^2}{4}$$

the region in which the I-th affine term is the unique maximum is exactly the open power cell of the site (p_I, ω_I). That is,

$$\langle q, v_I \rangle + b_I > \langle q, v_J \rangle + b_J \quad \Longleftrightarrow \quad \pi_I(q) < \pi_J(q).$$

Here, $\pi_I(q)$ denotes the power distance of the power-diagram site (p_I, ω_I) associated with the k-element support set I, defined as $\pi_I(q) := \|q - p_I\|^2 - \omega_I$.

Combining this with Eq. (4) gives

$$\mathcal{R}_I = \left\{ q \in \mathbb{R}^{d_k} : \pi_I(q) < \pi_J(q),\ \forall J \neq I,\ |J| = k \right\}.$$

Therefore, the strict Top-k routing region is exactly the interior of the corresponding power cell, and all nonempty power cells form a power diagram indexed by k-element support sets. □

B.3 Proposition 3.4 (Hypersimplex Representation of the Top-k Newton Polytope)

Proof. For any k-element subset $I \subseteq [N]$, its indicator vector can be written as

$$\mathbf{1}_I = \sum_{i \in I} e_i.$$

The linear map $K^\top e_i = k_i$ gives

$$K^\top \mathbf{1}_I = \sum_{i \in I} K^\top e_i = \sum_{i \in I} k_i = v_I.$$

Using the fact that a linear map commutes with the convex-hull operation,

$$\begin{aligned} K^\top \Delta_{k,N} &= K^\top \operatorname{conv}\{\mathbf{1}_I : |I| = k\} \\ &= \operatorname{conv}\left\{K^\top \mathbf{1}_I : |I| = k\right\} \\ &= \operatorname{conv}\left\{\sum_{i \in I} k_i : |I| = k\right\} = P_{N,k}. \end{aligned}$$

Therefore, the Top-k Newton polytope is the image of the (k, N)-hypersimplex under the linear map $K^\top$, and its routing geometry can equivalently be understood through the normal fan of this projected polytope. □

B.4 Proposition 3.6 (Local Power-Diagram Family Representation of Sliding-Window Attention)

Proof. By the definition of the banded mask,

$$M_{tj} = \begin{cases} 0, & j \in \mathcal{W}_t, \\ -\infty, & j \notin \mathcal{W}_t. \end{cases}$$

the local tropical polynomial of sliding-window attention at position t is

$$\begin{aligned} H_t(q_t) &:= \max_{j \in [N]} \left\{ \langle q_t, k_j \rangle + \beta_{tj} + M_{tj} \right\} \\ &= \max_{j \in \mathcal{W}_t} \left\{ \langle q_t, k_j \rangle + \beta_{tj} \right\}. \end{aligned} \tag{5}$$

Thus, key $j \in \mathcal{W}_t$ is selected by hard routing at position t if and only if

$$\langle q_t, k_j \rangle + \beta_{tj} > \langle q_t, k_\ell \rangle + \beta_{t\ell}, \quad \forall \ell \in \mathcal{W}_t \setminus \{j\}. \tag{6}$$

Eq. (5) is a max-affine function with slope k_j and bias β_{tj}. Set

$$p_{t,j} = \frac{k_j}{2}, \quad \omega_{t,j} = \beta_{tj} + \frac{\|k_j\|^2}{4}$$

The region in which the j-th affine term is the unique maximum is the open power cell of the weighted site $(p_{t,j}, \omega_{t,j})$. Thus, for any $j, \ell \in \mathcal{W}_t$,

$$\langle q_t, k_j \rangle + \beta_{tj} > \langle q_t, k_\ell \rangle + \beta_{t\ell} \quad \Longleftrightarrow \quad \pi_{t,j}(q_t) < \pi_{t,\ell}(q_t).$$

Combining this with Eq. (6) gives

$$\mathcal{R}_{t,j} = \left\{ q_t \in \mathbb{R}^d : \pi_{t,j}(q_t) < \pi_{t,\ell}(q_t),\ \forall \ell \in \mathcal{W}_t \setminus \{j\} \right\}.$$

Therefore, the closures of all strict hardmax routing regions at position t form the power diagram generated by the local site set $\mathcal{S}_t$.

Furthermore, as the position t changes, the window moves along the sequence and the corresponding local site set $\mathcal{S}_t$ changes with it. This gives the family of local power diagrams

$$\mathcal{D}_{\mathrm{SWA}} = \{\operatorname{Pow}(\mathcal{S}_t) : t = 1, \ldots, L\}.$$

If $q_1, \ldots, q_L$ are treated as independent variables, then for any joint routing label

$$\boldsymbol{j} = (j_1, \ldots, j_L), \quad j_t \in \mathcal{W}_t,$$

the corresponding joint strict routing region is

$$\mathcal{R}_{\boldsymbol{j}} = \{(q_1, \ldots, q_L) : q_t \in \mathcal{R}_{t,j_t},\ t = 1, \ldots, L\} = \prod_{t=1}^{L} \mathcal{R}_{t,j_t}.$$

Thus, the joint routing partition of the whole layer is the Cartesian product of the local routing partitions at all positions. □

B.5 Theorem 4.1 (Upper Bound on Top-k Routing Regions)

Proof. By Proposition 3.4, the Newton polytope corresponding to Top-k routing is

$$P_{N,k} = \mathrm{Newt}(F_k) = \mathrm{conv}\{v_I : |I| = k\}.$$

By tropical geometry, the full-dimensional linear regions induced by a tropical polynomial are given by the normal fan of its Newton polytope. For any support set I, its strict Top-k routing region satisfies

$$\mathcal{R}_I = \{q : \langle q, v_I \rangle + b_I > \langle q, v_J \rangle + b_J, \forall J \neq I\},$$

This region corresponds to the open normal cone of the Newton-polytope vertex v_I. Different nonempty strict routing regions correspond to different nonempty full-dimensional normal cones. Therefore,

$$R^{(1)}_{\mathrm{Top-k}} \leq |\,\mathrm{vert}(P_{N,k})|.$$

On the other hand,

$$P_{N,k} = \mathrm{conv}\left\{\sum_{i \in I} k_i : I \subseteq [N], |I| = k\right\}$$

has the following number of candidate generating points:

$$\binom{N}{k}.$$

Because the number of vertices of a polytope does not exceed the number of its generating points,

$$|\,\mathrm{vert}(P_{N,k})| \leq \binom{N}{k}.$$

In summary,

$$R^{(1)}_{\mathrm{Top-k}} \leq \binom{N}{k}.$$

Because the query variables are independent, for a joint routing label

$$\boldsymbol{I} = (I_1, \ldots, I_L), \quad |I_t| = k,$$

the corresponding joint region for the whole sequence is

$$\begin{aligned} \mathcal{R}_{\boldsymbol{I}} &= \{(q_1, \ldots, q_L) : q_t \in \mathcal{R}_{I_t},\ t = 1, \ldots, L\} \\ &= \mathcal{R}_{I_1} \times \cdots \times \mathcal{R}_{I_L}. \end{aligned} \tag{7}$$

Each position has at most $R^{(1)}_{\text{Top-}k}$ nonempty full-dimensional routing regions, so there are at most

$$\left(R^{(1)}_{\text{Top-}k}\right)^L$$

joint labels. Hence,

$$R_{\text{Top-}k,\mathrm{seq}} \leq \left(R^{(1)}_{\text{Top-}k}\right)^L.$$

Using the combinatorial label bound $R^{(1)}_{\text{Top-}k} \leq \binom{N}{k}$ gives

$$R_{\text{Top-}k,\mathrm{seq}} \leq \left(R^{(1)}_{\text{Top-}k}\right)^L \leq \left(\binom{N}{k}\right)^L.$$

Then we can get $R_{\text{Top-}k,\mathrm{seq}} \leq \left(R^{(1)}_{\text{Top-}k}\right)^L \leq \binom{N}{k}^L$. □

B.6 THEOREM 4.2 (EXPLICIT CONSTRUCTIVE LOWER BOUND FOR TOP-k ROUTING)

Proof. We want to keep the required dimension of the key and query spaces in the construction as small as possible. If $k \le N/2$, we only need to construct k largest scores and prove that they form the unique Top-k support set. Otherwise, we only need to construct $N-k$ smallest scores and prove that the remaining set is the unique Top-k support set.

Let

$$m := \min\{k, N-k\}, \tag{8}$$

Choose distinct real numbers $t_1 < \cdots < t_N$ and define the N keys k_i on the curve

$$\gamma(t) := (t, t^2, \ldots, t^{2m}) \in \mathbb{R}^{2m}, \quad k_i := \gamma(t_i). \tag{9}$$

If $k \le N/2$, then $m = k$. For any k-element subset I, define the polynomial

$$p_I(t) := -\prod_{i \in I} (t - t_i)^2 = a_0 + a_1 t + \cdots + a_{2k} t^{2k} \tag{10}$$

and the query $q_I = (a_1, \ldots, a_{2k})$. By Eq. (9),

$$\langle q_I, k_j \rangle = p_I(t_j) - a_0.$$

Clearly, if $j \in I$, then $p_I(t_j) = 0$; if $j \notin I$, then $p_I(t_j) < 0$. Therefore,

$$\langle q_I, k_i \rangle > \langle q_I, k_j \rangle, \quad \forall i \in I,\ j \notin I,$$

Thus, I is the unique Top-k support set.

Now consider $k > N/2$, so $m = N - k$. Let $I^c = [N] \setminus I$ and define

$$p_I(t) := \prod_{j \in I^c} (t - t_j)^2 = a_0 + a_1 t + \cdots + a_{2m} t^{2m}. \tag{11}$$

Take $q_I = (a_1, \ldots, a_{2m})$. Then

$$\langle q_I, k_j \rangle = p_I(t_j) - a_0.$$

For $j \in I^c$, we have $p_I(t_j) = 0$, while for $i \in I$, we have $p_I(t_i) > 0$. Therefore, I is again the unique Top-k support set.

In both constructions, the scores inside the support set are strictly larger than those outside it. By continuity, there is an open neighborhood of q_I in which this strict inequality remains true, so the corresponding region is full-dimensional and nonempty. All $\binom{N}{k}$ support sets can be realized. Combining this result with the combinatorial upper bound in Theorem 4.1 gives $\max R^{(1)}_{\text{Top-}k} = \binom{N}{k}$. The Cartesian product for independent queries gives $\max R_{\text{Top-}k,\text{seq}} = \binom{N}{k}^L$.. □

B.7 THEOREM 4.3 (UPPER BOUND ON SLIDING-WINDOW HARDMAX ROUTING REGIONS)

Proof. A hardmax region is labeled only by the winning index j in the window, so there are at most n_t candidate labels. For a fixed j, the feasible set

$$\begin{aligned} \mathcal{C}_{t,j} &:= \{q_t : s_{tj}(q_t) > s_{t\ell}(q_t), \quad \forall \ell \in \mathcal{W}_t \setminus \{j\}\} \\ &= \bigcap_{\ell \in \mathcal{W}_t \setminus \{j\}} \{q_t : \langle q_t, k_j - k_\ell \rangle > \beta_{t\ell} - \beta_{tj}\}. \end{aligned}$$

is convex and cannot split into several regions. Therefore,

$$R^{(1)}_{\text{SWA},t} \le n_t \le w. \tag{12}$$

For independent queries, the partition for the whole sequence is the Cartesian product of the local partitions at all positions. Thus,

$$R^{(1)}_{\text{SWA,seq}} \le \prod_{t=1}^{L} n_t \le w^L. \tag{13}$$

. □

B.8 Theorem 4.4 (Explicit Constructive Lower Bound for Sliding-Window Hardmax Routing)

Proof. Take the two-dimensional shared keys

$$\widetilde{k}_j := \begin{pmatrix} j \\ -j^2 \end{pmatrix}, \quad j = 1, \dots, N, \tag{14}$$

and let the t-th query be

$$q_t(\tau_t) := \begin{pmatrix} \tau_t \\ 1 \end{pmatrix}. \tag{15}$$

The corresponding score is

$$s_{tj}(\tau_t) = j\tau_t - j^2.$$

For any $j \in \mathcal{W}_t$, take $\tau_t = 2j$. Then

$$s_{tj}(2j) = j^2,$$

For any $\ell \neq j$,

$$s_{t\ell}(2j) = 2j\ell - \ell^2 = j^2 - (\ell - j)^2 < j^2.$$

Therefore, the score of the j-th key is the largest. The inequality remains true in a neighborhood of $\tau_t = 2j$, so every candidate key has a nonempty full-dimensional region. Because $\tau_1, \dots, \tau_N$ can be chosen independently, the Cartesian product of any collection of local routes is nonempty, and the product upper bound is attained. □

B.9 Theorem 4.6 (Equal-Budget Geometric-Capacity Comparison)

Proof. When $k = w = B$, both mechanisms retain B KV pairs at every query position. Therefore, by Definition 4.5, they have the same per-query post-routing KV activation-memory budget. By Theorems 4.1 and 4.2, under the given dimension condition, the maximum number of strict single-query routing regions for Top-B is $\binom{N}{B}$. By Theorems 4.3 and 4.4, the maximum number of strict single-query routing regions for a full sliding window of width B is B. Dividing the two numbers gives $\frac{R^{(1),\max}_{\text{Top-}B}}{R^{(1),\max}_{\text{SWA},t}} = \frac{\binom{N}{B}}{B}$

If $q_1, \dots, q_L$ can vary independently, then the joint routing partition is the Cartesian product of the single-query partitions. Therefore, the maximum sequence-level capacities of the two mechanisms are $\binom{N}{B}^L$ and B^L, respectively. Their ratio gives $\frac{R^{\max}_{\text{Top-}B,\text{seq}}}{R^{\max}_{\text{SWA,seq}}} = \left(\frac{\binom{N}{B}}{B}\right)^L$. □

C Related Work

Theoretical analysis of sliding-window attention. Sliding-window attention uses a banded mask to restrict the candidate keys of each query to a local neighborhood, so the amount of attention computation grows linearly with sequence length. Sparse Transformer and Longformer demonstrated the computational efficiency and task performance of local attention in long-sequence modeling through, respectively, sparse connection factorization and the combination of local windows with global connections (Child et al., 2019; Beltagy et al., 2020). On the theoretical side, Yun et al. (2020) gave the connectivity conditions required for sparse Transformers to retain universal approximation properties. Zaheer et al. (2020) proved that BigBird, which combines local, random, and global connections, has universal approximation properties and Turing completeness under certain conditions. It should be noted that BigBird's result depends on a combination of several types of sparse connections and does not concern the sliding-window structure alone. More recently, under a fixed-precision assumption, Li & Cotterell (2026) described the sequence operations that local attention can implement from the viewpoints of formal languages and linear temporal logic, and showed that local and global attention complement each other in processing temporal information. Overall, existing theories mainly treat the sliding window as a connection structure that limits the range of information flow and analyze its role in function approximation, sequence computation, and information propagation in the full network. When the key vectors are fixed and the query varies continuously, the object of study becomes the geometric partition induced by the local candidate scores.

Theoretical analysis of Top-k attention. Unlike sliding-window attention, which fixes candidate positions in advance, global Top-k attention dynamically selects k candidates according to the scores between the query and all keys. Gupta et al. (2021) implemented exact Top-k attention through blockwise computation and analyzed its memory cost and task performance. Later work further studied training consistency, indexer recall, and implementation conditions for native Top-k sparse attention in long-context settings (Xiu et al., 2025; DeepSeek-AI et al., 2025). On the theoretical side, Tzachristas et al. (2026) established upper bounds on the distribution and output errors of Top-k attention using the truncated attention tail mass and logit gaps. Desai et al. (2026) studied approximate attention that combines Top-k selection with random sampling and gave corresponding probabilistic approximation guarantees. Other studies analyzed the sequence computations that Top-1 hard attention can perform from the viewpoints of formal languages and circuit complexity (Hahn, 2020; Bergsträßer et al., 2024). However, the single-key selection of Top-1 and the combinatorial support-set selection of Top-k have different discrete structures. Existing Top-k theories therefore mainly treat it as a truncation or approximation operator for dense attention and focus on how much attention mass is retained and how large the resulting output error is. If we instead examine how the selected support set changes with the query, the focus moves from approximation error to the region structure formed by combinatorial support sets in query space.

Maximum number of full-dimensional linear regions. For piecewise-linear neural networks, the input space can be divided into full-dimensional regions, and the network corresponds to a fixed affine map within each region. Therefore, the maximum number of full-dimensional linear regions that a given network architecture can produce is an important measure of the complexity of its geometric partition. Pascanu et al. (2013) studied the response regions of deep piecewise-linear networks from the viewpoint of computational geometry. By comparing the region counts of deep and shallow networks, they showed that network depth can greatly increase the complexity of the input-space partition. Montúfar et al. (2014) established combinatorial lower bounds on the number of linear regions for ReLU and Maxout networks and showed that the region count grows exponentially with network depth. Serra et al. (2018) further gave tighter upper and lower bounds for ReLU networks, exact results for one-dimensional inputs, and a region-enumeration method based on mixed-integer programming. Hanin & Rolnick (2019) studied typical region counts and boundary densities in randomly initialized and trained networks, and pointed out that the region complexity of practical networks may be much smaller than the maximum allowed by the architecture. Xiong et al. (2024) extended region counting to convolutional neural networks with general piecewise-linear activations, derived the maximum and average numbers of linear regions for one layer and upper and lower bounds for multiple layers, and showed that convolutional networks can form more regions per parameter than fully connected piecewise-linear networks. These studies established the basic definition, counting methods, and asymptotic analysis framework for the maximum number of full-dimensional linear regions. They also provide a common scale for comparing the geometric partitions induced by different network mechanisms. Unlike ReLU regions, which are determined by neuron activation states, the routing regions of sparse attention are determined by the order relations among several affine scores. We therefore need to describe the polyhedral structure formed by max-affine functions and their sets of tied maxima.

Applications of tropical geometry. Tropical geometry uses maximum and addition as its basic operations and provides a natural algebraic and geometric language for max-affine functions and the polyhedral partitions they induce. Charisopoulos & Maragos (2018) represented network layers with ReLU, Leaky ReLU, and Maxout activations as tropical polynomials and used this representation to derive upper bounds on the corresponding numbers of linear regions. Zhang et al. (2018) proved that ReLU networks can be represented as tropical rational maps and connected their linear regions and decision boundaries to the vertices of related polytopes and tropical hypersurfaces, respectively. Alfarra et al. (2022) further proved that the decision boundaries of certain piecewise-linear classification networks are contained in tropical hypersurfaces, and used the convex geometry of the related zonotopes to analyze network pruning and adversarial perturbations. Smyrnis & Maragos (2020) studied multiclass neural-network compression by approximating the Newton polytopes of tropical polynomials, applying tropical geometry to model simplification rather than only network-structure analysis. Pham & Garg (2024) represented message-passing graph neural networks with ReLU activations as tropical rational signomial maps and gave upper and lower bounds on the number of linear regions under different aggregation and update structures. Hashemi et al. (2026) constructed tropical attention operators directly over the max-plus semiring so that they can simulate combina-

torial algorithms based on dynamic programming. That work focuses on designing new attention operators rather than on the routing partitions of standard sparse attention.

Recent studies have begun to apply tropical geometry to conditional routing and Transformer attention mechanisms. Su et al. (2026) represented Top-k MoE routing as the k-th elementary symmetric tropical polynomial and connected the partition of expert support sets to the normal fan of a hypersimplex. Su & Liu (2026) proved that, with fixed keys and in the zero-temperature limit, Top-1 self-attention induces a power diagram in query space, and analyzed the region combinations produced by multi-head attention through Minkowski sums of Newton polytopes. These studies extend the use of tropical geometry from general piecewise-linear networks to specific routing and attention mechanisms, and they provide a direct basis for the region counting in this paper.