WHEN IS AN ATTENTION SINK ROBUST? A CONVEX-GEOMETRIC THEORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention sinks are token positions that receive disproportionately large attention across queries. Yet it remains unclear when a sink is robust, whether it is arbitrarily amplifiable, and how key perturbations and context growth change the required query norm. We answer these questions for a fixed, pretrained attention head under deterministic product uncertainty with compact convex key sets. We prove that the queries at which the candidate token is a robust attention sink at a prescribed confidence level form a closed convex region. Our first main theorem gives an exact entropy–distance formula for the minimum query norm needed to make the candidate token a robust attention sink. The distance term measures the separation between the sink and weighted combinations of competing-token keys, while the entropy term captures their collective softmax contribution. The theorem also determines exactly when the robust attention weight of candidate token can be amplified arbitrarily close to one. Our second main theorem determines how the minimum query norm needed to preserve the candidate token as a robust attention sink changes under key perturbations and context growth. Let ρ be the key-error radius, m the number of added competing tokens, and γ_0 the nominal separation from the sink key to the aggregate competing-key set. For fixed ρ, the required norm grows as $\Theta(\log m)$; for fixed m, it grows as $\Theta((\gamma_0 - 2\rho)^{-1})$ as ρ approaches $\gamma_0/2$, and no finite query norm suffices when $2\rho \geq \gamma_0$. Together, our results distinguish whether a robust sink is geometrically possible from how large a query norm is needed to realize it, providing a principled basis for auditing sink preservation under key errors and long-context inference.

1 INTRODUCTION

Softmax attention assigns each visible token a probability according to its query–key score relative to all competing scores (Vaswani et al., 2017). In trained language models, particular token positions, especially initial positions, often receive disproportionately large attention across queries and are known as *attention sinks* (Xiao et al., 2024). Sink behavior has been connected to streaming inference, massive activations, internal information processing, and post-training quantization (Sun et al., 2024; Son et al., 2024). Because attention sinks affect both model analysis and inference design, understanding when this behavior persists is important.

Prior work has studied the emergence and internal mechanisms of attention sinks (Gu et al., 2025; Guo et al., 2024; Ran-Milo et al., 2026), their possible computational roles (Barbero et al., 2025; Ran-Milo, 2026), and methods that suppress or reproduce them (Qiu et al., 2026; Zuhri et al., 2026; Son et al., 2024). Together, these studies explain important aspects of how sinks arise, what functions they may serve, and how they can be controlled. They leave a different post-training question unresolved: *whether sink behavior identified in a trained attention head persists when its keys or visible context change during inference.*

This question arises directly in two common inference settings. KV-cache quantization perturbs the key vectors used to compute attention (Liu et al., 2024; Hooper et al., 2024), while context growth introduces new competing tokens and additional terms into the softmax denominator (Nakanishi, 2025). Since every attention weight depends on all visible query–key scores, either change can alter which token receives most of the attention. Consequently, observing a sink on a collection of prompts does not determine **whether the same token remains a sink** after key errors or context

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.

extension. This distinction matters whenever sink behavior is used to interpret a trained head or to design an inference method that preserves, removes, or constructs sink tokens. Therefore, understanding the robustness of attention sinks under these changes is essential.

In this work, **robustness** asks whether sink behavior survives such changes. For a fixed pretrained attention head, candidate sink token, and query, we allow the sink and competing-token keys to vary within specified uncertainty sets. Sink behavior is robust at that query if the candidate retains a prescribed attention level for every admissible key realization. We study the following question:

Given a fixed, pretrained attention head and deterministic key uncertainty, for which queries does a designated token remain a robust attention sink, when can this attention be amplified arbitrarily by increasing the query norm, and how do key perturbations and context growth change the required query norm?

We address this question through **convex geometry**. Because query–key scores are linear, their worst-case values depend only on the convex hulls of these sets and are represented exactly by support functions. Robust sink formation thus becomes a convex-set separation problem: query direction determines whether separation is possible, while query norm scales the score differences. The entropy representation of log-sum-exp accounts for the joint softmax contribution of all competing tokens. Our contributions are as follows.

- **Exact characterization of robust sink queries.** Let A denote the set of admissible keys for the candidate sink, and let $B_1, \dots, B_N$ denote those of the $N \geq 1$ competing tokens. For an error tolerance $\varepsilon \in (0, 1/2)$, we characterize exactly the robust sink region $\mathcal{S}_\varepsilon$ of queries whose worst-case sink attention, defined as the minimum attention weight assigned to the candidate token over all admissible realizations of the sink and competing-token keys, is at least $1 - \varepsilon$. We prove that this region is closed and convex and identify exactly which query directions can reach it (Section 3).
- **Critical query norm and arbitrary sink amplification.** Our first main theorem gives an exact entropy–distance variational formula for the minimum query norm required to guarantee the prescribed worst-case sink attention. The entropy term captures how multiple competing tokens contribute collectively to the softmax denominator, while the distance term measures how effectively their key mixture can be separated from the sink uncertainty set. The theorem also proves that the robust sink is arbitrarily amplifiable if and only if A has strictly positive distance from $\mathrm{conv}(\bigcup_{i=1}^{N} B_i)$ (Section 4).
- **Robustness phase transition under quantization and context growth.** Our second main theorem determines how bounded key errors and added tokens change the minimum query norm needed to preserve a robust sink. For a fixed post-positional key configuration, each key may vary within radius $\rho \geq 0$, and $m \geq 1$ added tokens share the same admissible key set. Let γ_0 be the distance from the nominal sink key to the aggregate nominal competing-key set. The critical query norm is finite exactly when $2\rho < \gamma_0$. Below this boundary, it grows logarithmically with m for fixed ρ and as $(\gamma_0 - 2\rho)^{-1}$ for fixed m as $\rho \uparrow \gamma_0/2$. Thus, context growth raises the required query norm, whereas sufficiently large key errors make robust sink preservation impossible at any finite norm (Section 5).

2 Problem Formulation

In this section, we formulate robust attention sinks for a fixed, pretrained attention head. We define the robust sink region under admissible key perturbations and introduce the query-norm bound used to study sink reachability and amplification. Appendix A reviews attention sinks and the required convex-analytic concepts, while Table 1 summarizes the notation.

In standard dot-product attention, queries and keys are vectors in $\mathbb{R}^d$, and their score is computed by the Euclidean inner product. We use a slightly more general normed-space formulation because different key-error models may be naturally measured by different norms, while the proofs depend only on linear query–key scores and norm duality. Readers interested in the Euclidean setting may take all pairings below to be ordinary dot products.

Formally, let $(V, |\cdot|)$ be a finite-dimensional real normed space of dimension $d := \dim V$, whose elements $k \in V$ are called *key vectors*. Let V^* be its dual space, whose elements $q \in V^*$ are linear

functionals on keys and are called *query vectors*. We write $\langle q, k\rangle$ for the resulting query–key score before temperature scaling and define the dual norm by $|q|_* := \sup_{|k|\le 1} |\langle q, k\rangle|$. Throughout the paper, we refer to $|q|_*$ as the *query norm*. For an original key $k \in V$ and its perturbed realization $k' \in V$, let $\Delta k := k' - k$ denote the key perturbation. The query norm controls the resulting change in the unscaled score: $|\langle q, \Delta k\rangle| \le |q|_* |\Delta k|$ In the standard Euclidean setting, $V = V^* = \mathbb{R}^d$, the pairing is the usual inner product, and the primal and dual norms coincide.

Assumption 2.1 (Fixed-head attention model)**.** We consider a single attention head whose parameters are fixed after training and condition on a fixed query position and configuration of visible token positions. All query and key projections and position-dependent transformations have already been applied; a different positional configuration defines a new instance of the framework. Let s denote a designated candidate sink token position, and let $N \ge 1$ be the number of other token positions visible to the current query. We call these positions the *competitors (competing tokens)* and index them by $[N] := \{1, \dots, N\}$. The resulting pre-softmax score between a query $q \in V^*$ and a key $k \in V$ is $\langle q, k\rangle/\tau$, where $\tau > 0$ is a fixed temperature.

Let $a \in V$ denote a realization of the key of token s, and let $b_i \in V$ denote a realization of the key of competitor $i \in [N]$. We write $b_{1:N} := (b_1, \dots, b_N) \in V^N$. Under Assumption 2.1, the attention weight assigned to token s is

$$p_s(q; a, b_{1:N}) := \frac{\exp(\langle q, a\rangle/\tau)}{\exp(\langle q, a\rangle/\tau) + \sum_{i=1}^{N} \exp(\langle q, b_i\rangle/\tau)}. \tag{1}$$

For standard scaled dot-product attention, $\tau = \sqrt{d}$ (Vaswani et al., 2017).

Assumption 2.2 (Product key uncertainty)**.** Let $A \subset V$ be the set of admissible realizations of the sink key, and let $B_i \subset V$ be the set of admissible realizations of competitor key $i \in [N]$. The sets $A, B_1, \dots, B_N$ are nonempty, compact, and convex, and the joint uncertainty set is $\mathcal{U} := A \times B_1 \times \cdots \times B_N$. Thus, no coupling constraints are imposed across tokens: each key may be selected adversarially within its own uncertainty set, simultaneously with all others.

Definition 2.3 (Robust attention weight)**.** For a query $q \in V^*$, the robust attention weight of token s is the smallest attention weight attained over the joint uncertainty set:

$$\underline{p}_s(q) := \min_{(a, b_{1:N}) \in \mathcal{U}} p_s(q; a, b_{1:N}). \tag{2}$$

Compactness of $\mathcal{U}$ and continuity of the attention weight ensure that the minimum in Eq. (2) is attained.

Remark 2.4 (Convexification invariance). Convexity in Assumption 2.2 entails no loss for product uncertainty sets. Indeed, Eq. (1) implies

$$\underline{p}_s(q) = \left[1 + \sum_{i=1}^{N} \exp\left(\frac{\sup_{b \in B_i} \langle q, b\rangle - \inf_{a \in A} \langle q, a\rangle}{\tau}\right)\right]^{-1}.$$

The extrema of a linear functional are unchanged when a compact set is replaced by its convex hull. Hence, any compact key uncertainty sets can be convexified without changing $\underline{p}_s(q)$.

Definition 2.5 (Key-side robustness and robust sink region)**.** Fix an error tolerance $\varepsilon \in (0, 1/2)$. Sink behavior is *ε-robust* at a query $q \in V^*$ if the candidate retains attention at least $1 - \varepsilon$ under every admissible key realization: $p_s(q; a, b_{1:N}) \ge 1 - \varepsilon$ for every $(a, b_{1:N}) \in \mathcal{U}$. Equivalently, $\underline{p}_s(q) \ge 1 - \varepsilon$. The corresponding robust sink region is

$$\mathcal{S}_\varepsilon := \{q \in V^* : \underline{p}_s(q) \ge 1 - \varepsilon\}. \tag{3}$$

Token s is called an ε-robust attention sink at every query $q \in \mathcal{S}_\varepsilon$.

The set in Eq. (3) is the primary geometric object of this work. It contains all queries that preserve the prescribed sink confidence uniformly over the admissible key perturbations.

Remark 2.6 (High-confidence regime). The restriction $\varepsilon \in (0, 1/2)$ ensures that the target attention level $1 - \varepsilon$ exceeds one half. Thus, under every admissible key realization, the candidate sink receives more attention than all competitors collectively. We use *robust attention sink* to refer to this worst-case, high-confidence regime. The restriction $\varepsilon < 1/2$ also underlies the strict-separation and outward-radial properties established in Section 3.

Definition 2.7 (Norm-constrained sink strength and critical query norm)**.** For a query-norm bound $R \geq 0$, define the query ball by $\mathcal{Q}_R := \{q \in V^* : |q|_* \leq R\}$. The *optimal robust sink probability* at norm bound R, denoted by P_R^*, is the largest robust attention weight attainable over $\mathcal{Q}_R$. The *critical query norm* for tolerance ε, denoted by $R_{\mathrm{crit}}(\varepsilon)$, is the smallest query-norm bound for which $\mathcal{Q}_R$ intersects $\mathcal{S}_\varepsilon$. Formally,

$$P_R^* := \max_{q \in \mathcal{Q}_R} \underline{p}_s(q), \quad R_{\mathrm{crit}}(\varepsilon) := \inf\{R \geq 0 : \mathcal{S}_\varepsilon \cap \mathcal{Q}_R \neq \varnothing\}. \tag{4}$$

Here, $\inf \varnothing := +\infty$. The candidate token s is *arbitrarily amplifiable* if $\lim_{R\to\infty} P_R^* = 1$.

The representation in Remark 2.4 and the compactness of $A, B_1, \ldots, B_N$ imply that $\underline{p}_s$ is continuous on V^*. Since V^* is finite dimensional, $\mathcal{Q}_R$ is compact, so the maximum defining P_R^* is attained for every $R \geq 0$. Thus, the two quantities in Eq. (4) capture complementary aspects of sink reachability: P_R^* is the largest worst-case sink probability attainable under the constraint $|q|_* \leq R$, whereas $R_{\mathrm{crit}}(\varepsilon)$ is the infimum query norm required to guarantee sink attention at least $1 - \varepsilon$, equivalently the dual-norm distance from the origin to $\mathcal{S}_\varepsilon$.

Remark 2.8 (Scope). The framework describes deterministic key-side robustness for a fixed attention head and a fixed positional snapshot. Norm-bounded quantization and cache errors are represented by the sets A and B_i, while context growth is represented by adding competitor sets. The query ball $\mathcal{Q}_R$ is an ambient capability set in V^*. If $\mathcal{M}_Q \subseteq V^*$ denotes the set of queries realizable by the model, define $R_{\mathrm{crit}}^{\mathcal{M}_Q}(\varepsilon) := \inf\{R \geq 0 : \mathcal{S}_\varepsilon \cap \mathcal{Q}_R \cap \mathcal{M}_Q \neq \varnothing\}$. Then $R_{\mathrm{crit}}(\varepsilon) \leq R_{\mathrm{crit}}^{\mathcal{M}_Q}(\varepsilon)$. Thus, the ambient critical query norm derived in this paper is a lower bound on the query norm required under model-realizability constraints; its finiteness does not by itself guarantee that a natural input sequence realizes a robust sink query.

The preceding formulation allows arbitrary nonempty compact convex key uncertainty sets. To obtain explicit results for key perturbations and context growth, Section 5 considers the following specialization of the product uncertainty model.

Assumption 2.9 (Uniform key perturbations and exchangeable context growth)**.** Let $n \geq 1$ be the number of existing competitor tokens. Let $k_s \in V$ denote the nominal sink key, and let $k_1, \ldots, k_n \in V$ denote the nominal competitor keys. Define the closed unit ball of V by $\mathbb{B} := \{v \in V : \|v\| \leq 1\}$, and let $\rho \geq 0$ be the common key-perturbation radius. The perturbed sink and existing competitor uncertainty sets are $A_\rho := k_s + \rho\mathbb{B}, \quad B_{i,\rho} := k_i + \rho\mathbb{B} \quad \text{for } i \in [n]$. Let $C \subset V$ be a nonempty compact convex set describing the nominal keys of newly added context tokens. For an integer $m \geq 1$, each of the m new competitors has uncertainty set $C_\rho := C + \rho\mathbb{B}$. The resulting attention problem has $n + m$ competitors.

3 Convex Geometry of the Robust Sink Region

In this section, we represent the robust sink region as a convex sublevel set based on the relative sink and competing-token keys, and derive its geometry and an exact criterion for feasible query directions.

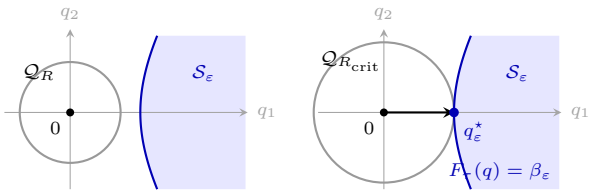
For an admissible sink key $a \in A$ and competitor key $b_i \in B_i$, their unscaled score difference is $\langle q, b_i - a\rangle$. We therefore define the *relative-key set* $D_i := B_i - A := \{b - a : b \in B_i,\ a \in A\}, \quad i \in [N]$. Each $d \in D_i$ is a possible competitor-minus-sink key difference. If $\langle q, d\rangle > 0$, the competitor scores higher than the sink for that realization; if it is negative, the sink scores higher. Under Assumption 2.2, each D_i is nonempty, compact, and convex.

For any nonempty compact set $C \subset V$, write $h_C(q) := \max_{x\in C}\langle q, x\rangle$ for its support function. Thus, $h_{D_i}(q)$ is the worst-case score advantage of competitor i over the sink. In particular, $h_{D_i}(q) < 0$ means that the sink outscores competitor i for every admissible key realization, while $\exp(h_{D_i}(q)/\tau)$ is the largest possible ratio of their unnormalized softmax weights.

To summarize the collective competition from all N competitors, define the *robust log-partition*

$$F_\tau(q) := \tau \log\left(\sum_{i=1}^{N} \exp\left(\frac{h_{D_i}(q)}{\tau}\right)\right). \tag{5}$$

The quantity $F_\tau(q)$ summarizes the total worst-case competition against the sink: a smaller $F_\tau(q)$ means that, even under the worst-case admissible key perturbations, the competitors contribute less



(a) $R < R_{\text{crit}}(\varepsilon)$ **(b)** $R = R_{\text{crit}}(\varepsilon)$

Figure 1: Schematic reachability under a query-norm constraint in a two-dimensional Euclidean query space. Both panels show the same robust sink region $\mathcal{S}_\varepsilon$. **(a)** When $R < R_{\text{crit}}(\varepsilon)$, the query ball is disjoint from the robust sink region. **(b)** At $R = R_{\text{crit}}(\varepsilon)$, the query ball first reaches the region at a minimum-norm query $q_\varepsilon^\star$. The blue curve is the target-confidence boundary $F_\tau(q) = \beta_\varepsilon$.

total softmax weight relative to the sink, so the sink receives more attention. A larger $F_\tau(q)$ means stronger collective competition and thus lower worst-case sink attention. The next two propositions use F_τ to characterize the robust sink region and determine which query directions can reach it.

Proposition 3.1 (Exact convex characterization)**.** *Define the sink threshold* $\beta_\varepsilon := \tau \log(\varepsilon/(1-\varepsilon)) < 0$. *Then, for every query* $q \in V^*$,

$$\underline{p}_s(q) = \frac{1}{1 + \exp(F_\tau(q)/\tau)}, \quad \mathcal{S}_\varepsilon = \{q \in V^* : F_\tau(q) \le \beta_\varepsilon\}. \tag{6}$$

The region $\mathcal{S}_\varepsilon$ *is closed and convex and does not contain the origin. It is also outward radial, meaning that it is closed under scaling away from the origin: for every* $q \in \mathcal{S}_\varepsilon$ *and every* $t \ge 1$, *we have* $tq \in \mathcal{S}_\varepsilon$. *This property relies on* $\varepsilon < 1/2$*: at any* $q \in \mathcal{S}_\varepsilon$, *the candidate sink scores strictly above every competing token under every admissible key realization, and scaling* q *enlarges all these score differences. If* $\mathcal{S}_\varepsilon$ *is nonempty, there exists a minimum-norm query* $q_\varepsilon^\star \in \mathcal{S}_\varepsilon$ *satisfying* $R_{\text{crit}}(\varepsilon) = \min_{q \in \mathcal{S}_\varepsilon} \|q\|_* = \|q_\varepsilon^\star\|_*, \quad F_\tau(q_\varepsilon^\star) = \beta_\varepsilon$.

Remark 3.2. Proposition 3.1 shows that the queries guaranteeing the prescribed robust sink confidence form the closed convex set $\mathcal{S}_\varepsilon$. Equivalently, a query belongs to this set exactly when $F_\tau(q) \le \beta_\varepsilon$, so robustness can be checked without examining every admissible key realization separately. Convexity means that interpolating between two robust queries remains robust. Closedness, together with the finite dimensionality of V^*, ensures that whenever $\mathcal{S}_\varepsilon$ is nonempty, it contains at least one query with the smallest norm. This norm is $R_{\text{crit}}(\varepsilon)$. Moreover, once a query is robust, increasing its norm without changing its direction preserves robustness. Thus, the proposition describes both the shape of the robust sink region and how it first becomes reachable as the allowed query norm increases. A detailed proof is given in Appendix B.

When $\mathcal{S}_\varepsilon$ is nonempty, the minimum-norm query $q_\varepsilon^\star$ determines the critical query norm $R_{\text{crit}}(\varepsilon)$. If $R < R_{\text{crit}}(\varepsilon)$, no query satisfying $|q|_* \le R$ can guarantee sink attention at least $1 - \varepsilon$ under all admissible key realizations. At $R = R_{\text{crit}}(\varepsilon)$, this guarantee first becomes achievable at $q_\varepsilon^\star$. It remains achievable for every larger norm bound because $\mathcal{Q}_R$ expands with R. Figure 1 compares the infeasible and first-feasible cases.

The nearest-point result determines how large the query norm must be before the robust sink region becomes reachable, but it does not identify ***which query directions can reach the region***. The next proposition answers this question using the worst-case sink-over-competitor score margin along each direction.

Proposition 3.3 (Directional sink criterion)**.** *Let* $u \in V^*$ *be a* unit query direction*, meaning that* $\|u\|_* = 1$. *Let* $r \in [0, \infty)$ *denote the query magnitude along* u*, so that the query* ru *satisfies* $\|ru\|_* = r$. *Define the query ray generated by* u *as* $\mathcal{R}_u := \{ru : r \in [0, \infty)\}$*, and define the* robust directional margin *of* u *by* $\delta(u) := -\max_{i \in [N]} h_{D_i}(u) = \min_{i \in [N]} \min_{\substack{a \in A \\ b \in B_i}} \langle u, a - b \rangle$. *Thus,* $\delta(u)$ *is the smallest sink-over-competitor score margin along* u*, taken over every competitor and every admissible key realization. The ray* $\mathcal{R}_u$ *intersects* $\mathcal{S}_\varepsilon$ *if and only if* $\delta(u) > 0$. *If* $\delta(u) > 0$*, then, for every query magnitude* $r \ge 0$,

$$\frac{1}{1 + N \exp(-r\delta(u)/\tau)} \le \underline{p}_s(ru) \le \frac{1}{1 + \exp(-r\delta(u)/\tau)}. \tag{7}$$

Consequently, $\underline{p}_s(ru) \to 1$ *as* $r \to \infty$. *In addition, the sufficient condition* $r \geq \tau\delta(u)^{-1}\log(N(1-\varepsilon)/\varepsilon)$ *guarantees that* $ru \in \mathcal{S}_\varepsilon$.

Remark 3.4. Proposition 3.3 separates the choice of a query direction from the query norm along that direction. The margin $\delta(u)$ is the sink's smallest score lead over all competitors and all admissible key realizations. If $\delta(u) > 0$, the sink is always ahead along direction u, and increasing r enlarges every such score gap. The competitors' softmax contributions relative to the sink then decrease exponentially, so the candidate's robust attention weight eventually reaches any prescribed confidence level below one. If $\delta(u) \leq 0$, some competitor can tie or outscore the sink under an admissible realization, and increasing the query norm along u cannot make the candidate a robust attention sink. Eq. (7) also shows that a larger margin requires less query norm, while more competitors increase the sufficient norm through the factor $\log N$. A detailed proof is given in Appendix C.

4 Entropy–Distance Duality and Robust Sink Amplification

In this section, we build on the geometric characterization of $\mathcal{S}_\varepsilon$ from Section 3 to determine how much query norm is required to make the candidate token a robust attention sink and when its robust attention weight can be amplified arbitrarily close to one. By Proposition 3.1, $\mathcal{S}_\varepsilon = \{q \in V^* : F_\tau(q) \leq \beta_\varepsilon\}$. These questions therefore reduce to finding the smallest value of F_τ attainable within each query ball $\mathcal{Q}_R$.

For a query-norm bound $R \geq 0$, write $\Psi_R := \min_{q\in\mathcal{Q}_R} F_\tau(q)$. The minimum is attained because $\mathcal{Q}_R$ is compact and F_τ is continuous. Since $x \mapsto (1+\exp(x/\tau))^{-1}$ is strictly decreasing, Eq. (6) gives

$$P_R^* = \frac{1}{1+\exp(\Psi_R/\tau)}. \tag{8}$$

Moreover, $\mathcal{S}_\varepsilon \cap \mathcal{Q}_R \neq \varnothing$ if and only if $\Psi_R \leq \beta_\varepsilon$. Thus, Ψ_R determines both the maximum robust attention weight attainable by the candidate token under the norm bound and whether that bound is sufficient to reach the robust sink region.

To evaluate Ψ_R, we use an auxiliary mixture variable describing how the competitors contribute collectively to the log-sum-exp. Let $\Delta_N := \{\pi = (\pi_1,\dots,\pi_N) \in \mathbb{R}_+^N : \sum_{i=1}^N \pi_i = 1\}$ be the probability simplex, and let $H(\pi) := -\sum_{i=1}^N \pi_i \log\pi_i$, $\quad 0\log 0 := 0$, be its Shannon entropy. For $\pi \in \Delta_N$, define the corresponding competitor mixture set by $B_\pi := \sum_{i=1}^N \pi_i B_i := \left\{\sum_{i=1}^N \pi_i b_i : b_i \in B_i \text{ for every } i \in [N]\right\}$. and write $d(\pi) := \operatorname{dist}(A, B_\pi)$. The entropy $H(\pi)$ is larger when the mixture spreads its weight across more competitors, while $d(\pi)$ measures how close the resulting collective competitor key can come to the sink uncertainty set. These quantities lead to the following dual characterization.

Theorem 4.1 (Entropy–distance duality)**.** *Under Assumptions 2.1 and 2.2, for every query-norm bound* $R \geq 0$,

$$\Psi_R = \max_{\pi\in\Delta_N} \{\tau H(\pi) - R\, d(\pi)\} = \max_{\pi\in\Delta_N} \{\tau H(\pi) - R\, \operatorname{dist}(A, B_\pi)\}. \tag{9}$$

The maximum is attained at a unique mixture $\pi_R^\star \in \Delta_N$.

Remark 4.2. Theorem 4.1 explains how competitor multiplicity and key separation jointly affect the candidate's robust attention weight. Because P_R^* decreases as Ψ_R increases, a larger dual objective means lower optimal worst-case sink attention. The entropy $H(\pi)$ is larger when the mixture weight is spread across more competitors, corresponding to several tokens contributing simultaneously to the softmax denominator and reducing the sink attention. Thus, adding identical or similar competitors can increase the required query norm even when the aggregate competing-key geometry is unchanged. The distance $d(\pi)$ has the opposite effect: a small distance means that the collective competitor key can lie close to the sink and is therefore difficult to suppress, whereas a large distance allows the query to create a larger score gap. The unique maximizer $\pi_R^\star$ balances these two effects and identifies the collective competitor pattern that is hardest to suppress at norm bound R, which need not be the strongest individual competitor. A detailed proof is given in Appendix D.

We next determine the query norm required to guarantee the prescribed sink confidence. Define $\lambda_\varepsilon := (1-\varepsilon)/\varepsilon > 1$, so that $\beta_\varepsilon = -\tau\log\lambda_\varepsilon$. The robust sink region is reachable under norm bound R exactly when $\Psi_R \leq -\tau\log\lambda_\varepsilon$. Substituting Eq. (9) gives the following exact threshold.

Corollary 4.3 (Exact critical query norm)**.** *The minimum query-norm bound required to attain robust sink probability at least* $1-\varepsilon$ *is*

$$R_{\mathrm{crit}}(\varepsilon) = \tau \max_{\pi \in \Delta_N} \frac{H(\pi) + \log \lambda_\varepsilon}{d(\pi)}. \tag{10}$$

The ratio in Eq. (10) *is defined as* $+\infty$ *whenever* $d(\pi) = 0$*. The maximum is attained.*

Remark 4.4. Each competitor mixture in Corollary 4.3 imposes a query-norm requirement. This requirement is larger when many competitors contribute through the softmax or when their collective key lies close to the sink. The critical query norm is the largest of these requirements and is therefore determined by the mixture that is hardest to suppress. If $d(\pi) = 0$ for some π, that mixture can reach the sink uncertainty set, so no finite query norm can guarantee attention at least $1-\varepsilon$. A detailed proof is given in Appendix E.

The critical-query-norm identifies zero mixture distance as the obstruction to finite-confidence reachability. To express this obstruction directly in terms of the original key sets, define the *aggregate competitor set* $B := \mathrm{conv}\left(\bigcup_{i=1}^N B_i\right)$ and its separation from the sink uncertainty set by $\gamma := \mathrm{dist}(A, B)$. The mixture bodies satisfy $B = \bigcup_{\pi \in \Delta_N} B_\pi$, so γ is the smallest value attained by $d(\pi)$. It follows that the same distance governing finite-norm separation also determines whether the sink can be amplified without bound.

Corollary 4.5 (Arbitrary amplification criterion)**.** *For every query-norm bound* $R \geq 0$*,*

$$\frac{1}{1 + N\exp(-R\gamma/\tau)} \leq P_R^* \leq \frac{1}{1+\exp(-R\gamma/\tau)}. \tag{11}$$

Consequently, $\lim_{R\to\infty} P_R^* = 1 \iff \gamma > 0 \iff A \cap B = \varnothing$*. Furthermore, the robust attention error has the asymptotic exponent* $\lim_{R\to\infty} -\frac{\tau}{R}\log(1-P_R^*) = \gamma$.

Remark 4.6. Corollary 4.5 gives a complete criterion for arbitrary robust amplification. If $\gamma > 0$, every collective competitor pattern remains separated from the sink, and increasing the query norm in an appropriate direction drives the worst-case sink attention toward one. If $\gamma = 0$, some competitor mixture can reach the sink uncertainty set, so increasing the query norm cannot remove this worst-case competition. Eq. (11) further shows that, when $\gamma > 0$, the robust attention error decreases exponentially with rate γ/τ. Thus, γ determines both whether arbitrary amplification is possible and how quickly it occurs. A detailed proof is given in Appendix F.

5 Quantization–Context Phase Diagram

In this section, we specialize the entropy–distance duality from Section 4 to the uniform key-perturbation and exchangeable context-growth model in Assumption 2.9. We derive the critical query norm required to preserve the candidate token as a robust attention sink, identify the perturbation threshold beyond which no finite query norm can provide this guarantee, and determine how the required norm scales with key error and context length.

To carry out this specialization, let $R_{\mathrm{crit}}(m, \rho; \varepsilon)$ denote the critical query norm from Definition 2.7. The general entropy–distance formula optimizes over mixtures of all $n+m$ competitors. Because the m added competitors share the same uncertainty set, their individual mixture weights can be summarized by their total weight without changing the corresponding mixture set.

Accordingly, define the reduced simplex $\mathcal{Z} := \left\{z = (\pi, \alpha) \in \mathbb{R}_+^n \times \mathbb{R}_+ : \sum_{i=1}^n \pi_i + \alpha = 1\right\}$, where $\pi = (\pi_1, \ldots, \pi_n)$ contains the mixture weights of the existing competitors and α is the total weight assigned to the m added competitors. For $z = (\pi, \alpha) \in \mathcal{Z}$, define the reduced entropy by $H_0(z) := -\sum_{i=1}^n \pi_i \log \pi_i - \alpha \log \alpha, \quad 0 \log 0 := 0$, and define the nominal competitor mixture set and its distance from the nominal sink key by

$$K_z := \sum_{i=1}^n \pi_i k_i + \alpha C, \quad r(z) := \mathrm{dist}(\{k_s\}, K_z) = \min_{c \in C} \left\| k_s - \sum_{i=1}^n \pi_i k_i - \alpha c \right\|.$$

Here, $\alpha C := \{\alpha c : c \in C\}$, with the convention $0C := \{0\}$. Finally, define the minimum nominal separation and its minimizing set by $\gamma_0 := \min_{z \in \mathcal{Z}} r(z), \quad \mathcal{M}_{\gamma_0} := \{z \in \mathcal{Z} : r(z) = \gamma_0\}$.

For a fixed reduced mixture $z = (\pi, \alpha) \in \mathcal{Z}$, the largest entropy obtained by distributing the total weight α among the m exchangeable added competitors is $H_0(z)+\alpha \log m$. Meanwhile, perturbing both the sink and competitor keys within radius ρ changes the distance between the sink uncertainty set and the corresponding competitor mixture set to $\operatorname{dist}(A_\rho, K_z + \rho\mathbb{B}) = (r(z) - 2\rho)_+$. Thus, the reduced mixture preserves exactly the entropy and distance terms required by the duality in Corollary 4.3. For $x \in \mathbb{R}$, write $(x)_+ := \max\{x, 0\}$, and recall that $\lambda_\varepsilon := (1 - \varepsilon)/\varepsilon > 1$. Applying the duality yields the following exact formula for the critical query norm, from which the perturbation boundary and the two asymptotic regimes follow.

Theorem 5.1 (Quantization–context phase diagram)**.** *Under Assumptions 2.1 and 2.9, the following statements hold.*

(i) Exact critical query norm. *For every integer $m \geq 1$ and every $\rho \geq 0$,*

$$R_{\mathrm{crit}}(m, \rho; \varepsilon) = \tau \sup_{z=(\pi,\alpha)\in\mathcal{Z}} \frac{H_0(z) + \alpha \log m + \log \lambda_\varepsilon}{(r(z) - 2\rho)_+}, \tag{12}$$

where a positive numerator divided by zero is defined as $+\infty$.

(ii) Sharp perturbation boundary. *The minimum nominal separation satisfies $\gamma_0 = \operatorname{dist}(\{k_s\}, \operatorname{conv}(\{k_1, \ldots, k_n\} \cup C))$, and the critical query norm obeys*

$$R_{\mathrm{crit}}(m, \rho; \varepsilon) < \infty \quad \Longleftrightarrow \quad 2\rho < \gamma_0. \tag{13}$$

(iii) Context and perturbation asymptotics. *For every fixed $\rho \in [0, \gamma_0/2)$,*

$$\lim_{\substack{m\to\infty \\ m\in\mathbb{N}}} \frac{R_{\mathrm{crit}}(m, \rho; \varepsilon)}{\tau \log m} = \max_{z=(\pi,\alpha)\in\mathcal{Z}} \frac{\alpha}{r(z) - 2\rho}. \tag{14}$$

If $\gamma_0 > 0$ and $m \geq 1$ is fixed, define $c_m(\varepsilon) := \max_{z=(\pi,\alpha)\in\mathcal{M}_{\gamma_0}} \{H_0(z) + \alpha \log m + \log \lambda_\varepsilon\}$. Then $c_m(\varepsilon) > 0$ and

$$\lim_{\rho\uparrow\gamma_0/2} \frac{(\gamma_0 - 2\rho) R_{\mathrm{crit}}(m, \rho; \varepsilon)}{\tau} = c_m(\varepsilon), \tag{15}$$

or equivalently, $R_{\mathrm{crit}}(m, \rho; \varepsilon) \sim \tau c_m(\varepsilon)/(\gamma_0 - 2\rho)$ as $\rho \uparrow \gamma_0/2$.

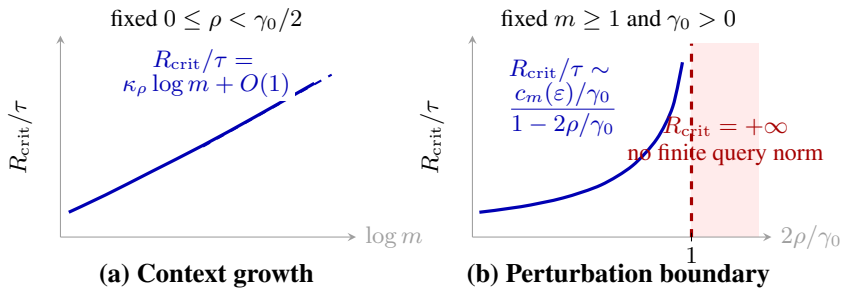
Remark 5.2. Theorem 5.1 shows that context growth and key perturbation affect robust sink formation in different ways. In Eq. (12), the m added competing tokens increase their collective contribution to the softmax denominator through the term $\alpha \log m$, while perturbations of both the sink and competing-token keys reduce their worst-case separation by 2ρ. Eq. (13) gives the exact error level beyond which no finite query norm can preserve the prescribed sink attention. Below this boundary, Eq. (14) shows that the required query norm grows logarithmically with the number of added tokens, whereas Eq. (15) shows that it diverges at an inverse-gap rate as the key error approaches the boundary. Thus, context growth raises the finite-temperature cost of suppressing competing tokens, while sufficiently large key perturbations make it impossible to preserve the candidate as a robust attention sink at any finite query norm. A detailed proof is given in Appendix G.

For fixed $\rho \in [0, \gamma_0/2)$, denote the context-growth coefficient in Eq. (14) by $\kappa_\rho := \max_{z=(\pi,\alpha)\in\mathcal{Z}} \frac{\alpha}{r(z)-2\rho}$, where $0 < \kappa_\rho < \infty$. In the perturbation limit, $\mathcal{M}_{\gamma_0}$ selects the competitor mixtures that lose separation first. Figure 2 illustrates these two limits.

6 Conclusion and Discussion

6.1 Conclusion

In this study, we characterized robust attention sinks in a fixed, pretrained attention head using convex geometry. We proved that the robust sink queries form a closed convex region and identified exactly which query directions reach it. Our entropy–distance duality gives the minimum query-norm bound and shows that arbitrary robust amplification is possible precisely when the sink uncertainty set is strictly separated from the aggregate competitor hull. Under norm-bounded key perturbations and exchangeable context growth, we established the feasibility boundary $2\rho = \gamma_0$, logarithmic query norm growth with the number of added tokens, and inverse-gap divergence near the boundary. More broadly, this analysis turns attention-sink robustness from a qualitative observation into a precise property of the key geometry, clarifying when sink behavior is structurally supported and when it becomes fragile under perturbation or longer contexts.



(a) Context growth **(b) Perturbation boundary**

Figure 2: Schematic one-parameter slices of the quantization–context phase diagram. **(a)** For fixed $\rho < \gamma_0/2$, the proof of Eq. (14) yields $R_{\mathrm{crit}}/\tau = \kappa_\rho \log m + O(1)$. **(b)** For fixed m and $\gamma_0 > 0$, Eq. (15) gives inverse-gap divergence as $2\rho/\gamma_0$ approaches one from below. The dashed vertical line marks the exact phase boundary in Eq. (13); at and beyond this boundary, $R_{\mathrm{crit}} = +\infty$.

6.2 Discussion

A sink is a relational property. Attention sinks are commonly identified by thresholding attention weights or averaging them over observed prompts. Our analysis indicates that this label is incomplete without specifying the attention head, query set, confidence level, and admissible key uncertainty. The same token can be a robust sink for one family of queries and unstable for another, while two tokens with similar nominal attention weights can have substantially different robustness margins. Sink behavior should therefore be understood as a relation among a query, a candidate key, and its competitors, rather than as an intrinsic property of a token or token position.

Implications for sink evaluation. Nominal attention weights, context length, and average quantization error are individually insufficient to predict sink stability. A large nominal weight need not persist under small key perturbations, and a long context need not be harmful when the added keys are well separated from the sink. The relevant quantities are instead the geometric separation from competitor mixtures and the query-norm bound available to overcome finite-temperature competition. This suggests evaluating trained heads through their robust sink regions, separation margins, and critical query norms, which distinguish genuinely robust sinks from attention concentration observed only on a particular prompt.

Practical use in head-level robustness audits. For a fixed head and prompt configuration, deterministic quantization or cache-error bounds can be used to construct the key uncertainty sets and to estimate γ_0 and $R_{\mathrm{crit}}(\varepsilon)$. The condition $2\rho \ge \gamma_0$ certifies that the prescribed sink confidence cannot be preserved at any finite query norm. More generally, if R_{arch} is an upper bound on the query norms realizable by the head, then $R_{\mathrm{crit}}(\varepsilon) > R_{\mathrm{arch}}$ rules out robust sink reachability. Within the feasible regime, the context-growth and perturbation laws quantify how much additional query norm is required as more tokens are added or key errors increase. These criteria can therefore support comparisons of quantization levels, cache configurations, and context-length limits when sink behavior is intended to be preserved or controlled.

Limitations and future work. The explicit quantization–context phase law assumes common norm-ball perturbations and exchangeable added competitors. Although the general entropy–distance duality allows heterogeneous compact convex sets, it retains product uncertainty and therefore excludes cross-token errors coupled by shared quantization parameters; under joint uncertainty, the tokenwise support-function decomposition and explicit phase laws need not persist. The analysis is also limited to attention probabilities in a single fixed head. Future work may address coupled perturbations and heterogeneous context growth, and extend the framework to value aggregation, multiple heads, residual connections, and cross-layer propagation.

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

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do llms attend to the first token? *arXiv preprint arXiv:2504.02732*, 2025.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004. ISBN 9780521833783. doi: 10.1017/CBO9780511804441. URL `https://web.stanford.edu/~boyd/cvxbook/`.

Nicola Cancedda. Spectral filters, dark signals, and attention sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4792–4808, 2024.

Thomas M Cover and Joy A Thomas. Elements of information theory, 2/e. 2006.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*, volume 2025, pp. 97114–97144, 2025.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I Jordan, and Song Mei. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in llms. *arXiv preprint arXiv:2410.13835*, 2024.

Jonghyun Hong and Sungyoon Lee. Variance sensitivity induces attention entropy collapse and instability in transformers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8371–8389, 2025.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W Mahoney, Yakun S Shao, Kurt Keutzer, and Amir Gholami. Kvquant: Towards 10 million context length llm inference with kv cache quantization. *Advances in Neural Information Processing Systems*, 37:1270–1303, 2024.

Jürgen Kindler. A simple proof of sion's minimax theorem. *The American Mathematical Monthly*, 112(4):356–358, 2005.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. *arXiv preprint arXiv:2402.02750*, 2024.

Ken M Nakanishi. Scalable-softmax is superior for attention. *arXiv preprint arXiv:2501.19399*, 2025.

Lawrence Narici and Edward Beckenstein. The hahn-banach theorem: the life and times. *Topology and its Applications*, 77(2):193–211, 1997.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *Advances in Neural Information Processing Systems*, 38:100092–100118, 2026.

Enrique Queipo-de Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in llms are two sides of the same coin. In *International Conference on Learning Representations*, volume 2026, pp. 13911–13937, 2026.

Yuval Ran-Milo. Attention sinks are provably necessary in softmax transformers: Evidence from trigger-conditional tasks. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 70–89, 2026.

Yuval Ran-Milo, Hila Ofek, and Shahar Mendel. A mechanistic account of attention sinks in gpt-2: One circuit, broader implications for mitigation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 90–98, 2026.

R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

R Tyrrell Rockafellar and Roger JB Wets. *Variational analysis*. Springer, 1998.

Walter Rudin. Real and complex analysis. 1974.

Valeria Ruscio, Umberto Nanni, and Fabrizio Silvestri. What are you sinking? a geometric approach on attention sink. *Advances in Neural Information Processing Systems*, 38:123482–123521, 2026.

Rolf Schneider. *Convex Bodies: The Brunn–Minkowski Theory*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2 edition, 2013.

Seungwoo Son, Wonpyo Park, Woohyun Han, Kyuyeun Kim, and Jaeho Lee. Prefixing attention sinks can mitigate activation outliers for large language model quantization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2242–2252, 2024.

A Stefănescu. A general min-max theorem. *Optimization*, 16(4):497–504, 1985.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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, 12 2008. ISSN 1935-8237. doi: 10.1561/2200000001. URL https://doi.org/10.1561/2200000001.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Stephen Zhang, Mustafa Khan, and Vardan Papyan. Attention sinks: A'catch, tag, release'mechanism for embeddings. *Advances in Neural Information Processing Systems*, 38: 83140–83181, 2026.

Zayd Muhammad Kawakibi Zuhri, Erland Hilman Fuadi, and Alham Fikri Aji. Softpick: No attention sink, no massive activations with rectified softmax. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 18179–18213, 2026.

A Attention Sink and Convex-Geometric Preliminaries

In this section, we review attention sinks and the convex-geometric and variational concepts used in the main analysis. We first describe the conventional meaning of an attention sink and then introduce the standard mathematical notation required by the proofs.

A.1 Attention Sinks

Let $(V, \|\cdot\|)$ be a finite-dimensional real normed space for keys, and let V^* be its dual query space. Given a query $q \in V^*$, visible keys $k_1, \dots, k_M \in V$, and a temperature $\tau > 0$, the attention probability assigned to token $j \in [M] := \{1, \dots, M\}$ is

$$p_j(q; k_{1:M}) := \frac{\exp(\langle q, k_j\rangle/\tau)}{\sum_{\ell=1}^{M} \exp(\langle q, k_\ell\rangle/\tau)},$$

where $\langle q, k_j\rangle$ is the dual pairing. When $V = \mathbb{R}^d$ is equipped with the Euclidean norm, standard scaled dot-product attention uses $\tau = \sqrt{d}$ (Vaswani et al., 2017).

An *attention sink* is a token that receives disproportionately large attention from a broad collection of queries, often despite having limited semantic relevance. Initial tokens are common sinks in autoregressive language models, although sink behavior may also occur at other token positions (Xiao et al., 2024; Gu et al., 2025). This conventional notion is based on observed attention patterns and does not specify a universal probability threshold or perturbation model. Definition 2.5 provides the formal notion used in this work by requiring a prescribed attention probability uniformly over admissible key realizations.

A.2 Convex-Geometric and Variational Tools

Dual norms. The dual norm associated with $\|\cdot\|$ is the standard operator norm on V^* (Rockafellar, 1997).

Definition A.1 (Dual norm and unit balls)**.** For $q \in V^*$, define $\|q\|_* := \sup_{\|x\|\le 1} |\langle q, x\rangle|$. The closed unit balls in V and V^* are

$$\mathbb{B} := \{x \in V : \|x\| \le 1\}, \quad \mathbb{B}_* := \{q \in V^* : \|q\|_* \le 1\}.$$

The dual norm satisfies $|\langle q, x\rangle| \le \|q\|_*\|x\|, \quad q \in V^*, \quad x \in V$. Because V is finite-dimensional, every closed and bounded subset of V or V^* is compact. In particular, the query ball $R\mathbb{B}_*$ is compact for every radius $R \ge 0$.

Convex hulls and Minkowski operations. We use the standard convex hull, Minkowski sum, and scalar multiplication of sets (Rockafellar, 1997; Schneider, 2013).

Definition A.2 (Convex hull and Minkowski operations)**.** For a nonempty set $C \subset V$, its convex hull is

$$\mathrm{conv}(C) := \left\{ \sum_{j=1}^{L} \theta_j x_j : L \ge 1,\ x_j \in C,\ \theta_j \ge 0,\ \sum_{j=1}^{L} \theta_j = 1 \right\}.$$

For nonempty sets $C, D \subset V$ and a scalar $t \ge 0$, define

$$C + D := \{c + d : c \in C,\ d \in D\}, \quad tC := \{tc : c \in C\}, \quad C - D := \{c - d : c \in C,\ d \in D\}.$$

Here, $C - D = C + (-D)$ is the Minkowski difference and is distinct from set-theoretic subtraction. Finite Minkowski sums and nonnegative scalar multiples preserve compactness and convexity. Moreover, if C is convex and $t_1, \dots, t_L \ge 0$, then

$$t_1 C + \cdots + t_L C = (t_1 + \cdots + t_L)C.$$

For compact convex sets $C_1, \dots, C_N \subset V$ and weights $\pi_1, \dots, \pi_N \ge 0$, the set $\sum_{i=1}^{N} \pi_i C_i$ is their weighted Minkowski sum.

Support functions. Support functions encode the extrema of linear functionals over sets and are standard objects in convex geometry (Schneider, 2013).

Definition A.3 (Support function)**.** For a nonempty compact set $C \subset V$, its support function $h_C : V^* \to \mathbb{R}$ is

$$h_C(q) := \max_{x \in C} \langle q, x \rangle.$$

The maximum is attained by compactness. The support function depends only on the convex hull of its argument and satisfies

$$h_{\mathrm{conv}(C)}(q) = h_C(q), \quad h_{C+D}(q) = h_C(q) + h_D(q), \quad h_{tC}(q) = t h_C(q)$$

for every $q \in V^*$ and $t \geq 0$. In particular,

$$h_{D-C}(q) = h_D(q) + h_C(-q) = \max_{d \in D} \langle q, d \rangle - \min_{c \in C} \langle q, c \rangle.$$

The function h_C is convex and positively homogeneous. If $M_C := \max_{x \in C} \|x\|$, then dual Hölder inequality gives

$$|h_C(q) - h_C(q')| \leq M_C \|q - q'\|_*, \quad q, q' \in V^*.$$

Thus, the support function of a compact set is Lipschitz continuous.

Distance and separation. Distances between sets and strict separation of compact convex sets are standard in convex and variational analysis (Rockafellar, 1997; Rockafellar & Wets, 1998).

Definition A.4 (Set distance)**.** For nonempty compact sets $C, D \subset V$, define

$$\mathrm{dist}(C, D) := \min_{c \in C,\ d \in D} \|c - d\|.$$

For $x \in V$, write $\mathrm{dist}(x, D) := \mathrm{dist}(\{x\}, D)$.

Compactness ensures that the minimum is attained and that

$$\mathrm{dist}(C, D) = 0 \quad \Longleftrightarrow \quad C \cap D \neq \varnothing.$$

If C and D are nonempty, compact, convex, and disjoint, the strong separation theorem provides a nonzero functional $q \in V^*$ such that

$$\sup_{c \in C} \langle q, c \rangle < \inf_{d \in D} \langle q, d \rangle.$$

This functional defines a direction that uniformly separates the two sets.

Probability simplex and Shannon entropy. Mixtures of competitors are indexed by the probability simplex, and their finite-temperature contribution is measured by Shannon entropy (Cover & Thomas, 2006).

Definition A.5 (Probability simplex and entropy)**.** For an integer $N \geq 1$, define

$$\Delta_N := \left\{ \pi = (\pi_1, \ldots, \pi_N) \in \mathbb{R}^N : \pi_i \geq 0,\ \sum_{i=1}^{N} \pi_i = 1 \right\}.$$

For $\pi \in \Delta_N$, its Shannon entropy is

$$H(\pi) := -\sum_{i=1}^{N} \pi_i \log \pi_i,$$

where $0 \log 0 := 0$ and all logarithms are natural.

The entropy function is continuous on Δ_N and satisfies $0 \leq H(\pi) \leq \log N$. The upper bound is attained by the uniform distribution, while the lower bound is attained at the vertices of Δ_N.

Log-sum-exp variational identity. The log-sum-exp function admits a variational representation in terms of Shannon entropy, also known as the Gibbs variational identity (Cover & Thomas, 2006; Wainwright & Jordan, 2008). For $x_1, \dots, x_N \in \mathbb{R}$ and $\tau > 0$,

$$\tau \log \left(\sum_{i=1}^{N} \exp(x_i/\tau) \right) = \max_{\pi \in \Delta_N} \left\{ \sum_{i=1}^{N} \pi_i x_i + \tau H(\pi) \right\}.$$

The unique maximizer is the softmax distribution

$$\pi_i^\star = \frac{\exp(x_i/\tau)}{\sum_{j=1}^{N} \exp(x_j/\tau)}, \quad i \in [N].$$

Minimax theorem. The interchange of query minimization and mixture maximization uses the finite-dimensional convex–concave form of Sion's minimax theorem (Stefănescu, 1985; Kindler, 2005). If X and Y are nonempty compact convex sets and $f : X \times Y \to \mathbb{R}$ is continuous, concave in its first argument, and convex in its second argument, then

$$\min_{y \in Y} \max_{x \in X} f(x, y) = \max_{x \in X} \min_{y \in Y} f(x, y).$$

Finally, for a scalar $t \in \mathbb{R}$, we use the positive-part notation $(t)_+ := \max\{t, 0\}$.

Table 1: Summary of the main notation. Competing tokens are all token positions visible to the current query other than the candidate sink.

Symbol	Meaning
Attention model and robustness	
V, V^*, $d = \dim V$	Finite-dimensional key space, its dual query space, and their dimension.
$k \in V$, $q \in V^*$	Key and query vectors.
$\langle q, k \rangle$, $\lvert k \rvert$, $\lvert q \rvert_*$	Unscaled query–key score, key norm, and query norm. Mathematically, $\lvert q \rvert_*$ is the dual norm induced by $\lvert \cdot \rvert$.
$\tau > 0$, $\mathbb{B}$	Softmax temperature and the closed unit ball $\mathbb{B} := \{v \in V : \lvert v \rvert \leq 1\}$.
s, N, $[N]$	Candidate sink position, number of competing tokens, and their index set $[N] := \{1, \dots, N\}$.
a, b_i, $b_{1:N}$	A realized sink key, the realized key of competing token i, and the tuple of all realized competing-token keys.
A, B_i, $\mathcal{U}$	Sets of admissible key vectors for the sink and competing token i, and the product uncertainty set $\mathcal{U} := A \times B_1 \times \cdots \times B_N$.
$p_s(q; a, b_{1:N})$	Attention weight assigned to the candidate sink for a fixed query and fixed key realizations.
$\underline{p}_s(q)$	Worst-case sink attention, obtained by minimizing p_s over all admissible key realizations in $\mathcal{U}$.
$\varepsilon \in (0, 1/2)$	Allowed attention error; the prescribed sink-attention level is $1 - \varepsilon$.
λ_ε, β_ε	Equivalent confidence parameters $\lambda_\varepsilon := (1 - \varepsilon)/\varepsilon$ and $\beta_\varepsilon := -\tau \log \lambda_\varepsilon$.
$\mathcal{S}_\varepsilon$	Robust sink region: queries guaranteeing sink attention at least $1 - \varepsilon$ for every admissible key realization.
R, $\mathcal{Q}_R$	Query-norm bound and query ball $\mathcal{Q}_R := \{q \in V^* : \lvert q \rvert_* \leq R\}$.
P_R^*	Largest worst-case sink attention attainable over $\mathcal{Q}_R$.
$R_{\mathrm{crit}}(\varepsilon)$	Minimum query norm required to reach $\mathcal{S}_\varepsilon$.
$\mathcal{M}_Q$	Set of queries realizable by the model.
$R_{\mathrm{crit}}^{\mathcal{M}_Q}(\varepsilon)$	Critical query norm after restricting queries to $\mathcal{M}_Q$.
R_{arch}	An upper bound on the query norms realizable by the attention head.
Geometry of the robust sink region	

Continued on the next page

Table 1: Summary of notation (continued)

Symbol	Meaning
$D_i := B_i - A$	Relative-key set containing all possible competing-token-minus-sink key differences for token i.
$h_C(q) := \max_{x \in C} \langle q, x \rangle$	Support function of a compact set C. In particular, $h_{D_i}(q)$ is the largest possible score advantage of competing token i over the sink.
$F_\tau(q)$	Robust log-partition combining the worst-case score advantages of all competing tokens through log-sum-exp.
$q_\varepsilon^\star$	A minimum-norm query in $\mathcal{S}_\varepsilon$, satisfying $\lvert q_\varepsilon^\star \rvert_* = R_{\mathrm{crit}}(\varepsilon)$.
u, t, $\mathcal{R}_u$	Unit query direction, query magnitude, and query ray $\mathcal{R}_u := \{tu : t \geq 0\}$.
$\delta(u)$	Smallest sink-over-competitor score margin along u, over all competing tokens and admissible key realizations.
Entropy–distance duality	
$\Psi_R := \min_{q \in \mathcal{Q}_R} F_\tau(q)$	Smallest robust log-partition attainable under query-norm bound R.
Δ_N, $\pi \in \Delta_N$	Probability simplex over the competing tokens and an auxiliary competitor-mixture weight vector.
$H(\pi)$	Shannon entropy of π, measuring how broadly the collective softmax competition is distributed across tokens.
$B_\pi := \sum_{i=1}^N \pi_i B_i$	Weighted Minkowski mixture of the competing-token key sets.
$d(\pi) := \mathrm{dist}(A, B_\pi)$	Distance between the admissible sink keys and the competitor mixture set B_π.
$\pi_R^\star$	Unique mixture maximizing the entropy–distance objective at query-norm bound R.
$B := \mathrm{conv}(\bigcup_{i=1}^N B_i)$	Aggregate convex set of all admissible competing-token keys.
$\gamma := \mathrm{dist}(A, B)$	Separation between the admissible sink keys and the aggregate competitor set in the general model.
Key perturbations and context growth	
n, m	Numbers of existing and newly added competing tokens.
k_s, k_i	Nominal sink key and nominal key of existing competing token i.
C	Compact convex set of nominal keys available to the added competing tokens.
$\rho \geq 0$	Common key-perturbation radius.
A_ρ, $B_{i,\rho}$, C_ρ	Perturbed key sets of the sink, existing competing token i, and the added competing tokens.
$z = (\pi, \alpha)$, $\mathcal{Z}$	Reduced competitor mixture and its simplex; α is the total mixture weight assigned to the m added tokens.
$H_0(z)$	Entropy of the reduced $(n+1)$-component mixture before distributing the weight α among the m added tokens.
$K_z := \sum_{i=1}^n \pi_i k_i + \alpha C$	Nominal competing-key mixture set associated with z.
$r(z) := \mathrm{dist}(\{k_s\}, K_z)$	Nominal distance between the sink key and the mixture set K_z.
$\gamma_0 := \min_{z \in \mathcal{Z}} r(z)$	Nominal separation between the sink key and the aggregate nominal competing-key set.
$\mathcal{M}_{\gamma_0}$	Set of reduced mixtures attaining the minimum nominal separation γ_0.
κ_ρ	Leading coefficient of the logarithmic growth of $R_{\mathrm{crit}}(m, \rho; \varepsilon)$ with m.
$c_m(\varepsilon)$	Coefficient of the inverse-gap divergence as $\rho \uparrow \gamma_0/2$.
Set operations	
$\mathrm{conv}(C)$	Convex hull of a set C.
$C + D$, αC	Minkowski sum and nonnegative scalar multiple of sets.
$\mathrm{dist}(C, D)$	Minimum distance between two nonempty compact sets C and D.
$(x)_+ := \max\{x, 0\}$	Positive part of a real number.

B PROOF OF PROPOSITION 3.1

Proof. We first establish the geometric properties of the relative-key bodies. Fix $i \in [N]$. Since A and B_i are nonempty, the set $D_i = B_i - A$ is nonempty. The product $B_i \times A$ is compact, and the map $\phi_i : B_i \times A \to V$ defined by $\phi_i(b, a) := b - a$ is continuous. Therefore, $D_i = \phi_i(B_i \times A)$ is compact.

To prove convexity, let $d^{(1)}, d^{(2)} \in D_i$ and $\theta \in [0, 1]$. By the definition of D_i, there exist $b^{(1)}, b^{(2)} \in B_i$ and $a^{(1)}, a^{(2)} \in A$ such that $d^{(1)} = b^{(1)} - a^{(1)}$ and $d^{(2)} = b^{(2)} - a^{(2)}$. Hence,

$$\begin{aligned}\theta d^{(1)} + (1-\theta)d^{(2)} &= \theta\big(b^{(1)} - a^{(1)}\big) + (1-\theta)\big(b^{(2)} - a^{(2)}\big) \\ &= \big(\theta b^{(1)} + (1-\theta)b^{(2)}\big) - \big(\theta a^{(1)} + (1-\theta)a^{(2)}\big).\end{aligned}$$

Convexity of B_i and A implies that $\theta b^{(1)} + (1-\theta)b^{(2)} \in B_i$ and $\theta a^{(1)} + (1-\theta)a^{(2)} \in A$. Therefore, $\theta d^{(1)} + (1-\theta)d^{(2)} \in D_i$, proving that D_i is convex.

We next compute the support function of D_i. For every $q \in V^*$,

$$\begin{aligned}h_{D_i}(q) &= \max_{d \in D_i} \langle q, d\rangle \\ &= \max_{\substack{a \in A \\ b \in B_i}} \langle q, b - a\rangle \\ &= \max_{\substack{a \in A \\ b \in B_i}} \big(\langle q, b\rangle - \langle q, a\rangle\big) \\ &= \max_{b \in B_i} \langle q, b\rangle - \min_{a \in A} \langle q, a\rangle.\end{aligned}$$

The maxima and minima above are attained because A and B_i are compact and the corresponding linear functionals are continuous.

Fix $q \in V^*$ and define $G_q : \mathcal{U} \to (0, \infty)$ by

$$G_q(a, b_{1:N}) := \sum_{i=1}^{N} \exp\left(\frac{\langle q, b_i\rangle - \langle q, a\rangle}{\tau}\right).$$

Dividing the numerator and denominator of Eq. (1) by $\exp(\langle q, a\rangle/\tau)$ gives

$$p_s(q; a, b_{1:N}) = \frac{1}{1 + G_q(a, b_{1:N})}.$$

Since $x \mapsto (1 + x)^{-1}$ is strictly decreasing on $[0, \infty)$ and G_q attains its maximum on the compact set $\mathcal{U}$,

$$\underline{p}_s(q) = \min_{(a, b_{1:N}) \in \mathcal{U}} \frac{1}{1 + G_q(a, b_{1:N})} = \frac{1}{1 + \max_{(a, b_{1:N}) \in \mathcal{U}} G_q(a, b_{1:N})}.$$

Define

$$\alpha(q) := \min_{a \in A} \langle q, a\rangle, \quad \mu_i(q) := \max_{b \in B_i} \langle q, b\rangle.$$

For every $(a, b_{1:N}) \in \mathcal{U}$, we have $\langle q, a\rangle \geq \alpha(q)$ and $\langle q, b_i\rangle \leq \mu_i(q)$. Therefore,

$$\begin{aligned}G_q(a, b_{1:N}) &= \exp\left(-\frac{\langle q, a\rangle}{\tau}\right) \sum_{i=1}^{N} \exp\left(\frac{\langle q, b_i\rangle}{\tau}\right) \\ &\leq \exp\left(-\frac{\alpha(q)}{\tau}\right) \sum_{i=1}^{N} \exp\left(\frac{\mu_i(q)}{\tau}\right).\end{aligned}$$

Compactness provides $a^\star \in A$ and $b_i^\star \in B_i$ satisfying $\langle q, a^\star \rangle = \alpha(q)$ and $\langle q, b_i^\star \rangle = \mu_i(q)$ for every $i \in [N]$. Since $(a^\star, b_1^\star, \dots, b_N^\star) \in \mathcal{U}$, the preceding upper bound is attained. Consequently,

$$\begin{aligned}\max_{(a,b_{1:N})\in\mathcal{U}} G_q(a, b_{1:N}) &= \exp\left(-\frac{\alpha(q)}{\tau}\right) \sum_{i=1}^{N} \exp\left(\frac{\mu_i(q)}{\tau}\right) \\ &= \sum_{i=1}^{N} \exp\left(\frac{\mu_i(q) - \alpha(q)}{\tau}\right) \\ &= \sum_{i=1}^{N} \exp\left(\frac{h_{D_i}(q)}{\tau}\right) \\ &= \exp\left(\frac{F_\tau(q)}{\tau}\right).\end{aligned}$$

It follows that

$$\underline{p}_s(q) = \frac{1}{1 + \exp(F_\tau(q)/\tau)}.$$

We now derive the sublevel-set representation. Since $\tau > 0$, $\varepsilon \in (0, 1)$, and the exponential function is positive, each of the following transformations is equivalent:

$$\begin{aligned}\underline{p}_s(q) \geq 1 - \varepsilon &\iff \frac{1}{1 + \exp(F_\tau(q)/\tau)} \geq 1 - \varepsilon \\ &\iff 1 \geq (1 - \varepsilon)\left(1 + \exp(F_\tau(q)/\tau)\right) \\ &\iff \varepsilon \geq (1 - \varepsilon)\exp(F_\tau(q)/\tau) \\ &\iff \exp(F_\tau(q)/\tau) \leq \frac{\varepsilon}{1 - \varepsilon} \\ &\iff F_\tau(q) \leq \tau \log\left(\frac{\varepsilon}{1 - \varepsilon}\right) \\ &\iff F_\tau(q) \leq \beta_\varepsilon.\end{aligned}$$

Thus,

$$\mathcal{S}_\varepsilon = \{q \in V^* : F_\tau(q) \leq \beta_\varepsilon\}.$$

We next prove that F_τ is continuous and convex. Let $C \subset V$ be nonempty and compact, and define $M_C := \max_{x\in C} \|x\| < \infty$. For any $q, q' \in V^*$,

$$\begin{aligned}h_C(q) &= \max_{x\in C}\big(\langle q', x\rangle + \langle q - q', x\rangle\big) \\ &\leq h_C(q') + \max_{x\in C} |\langle q - q', x\rangle| \\ &\leq h_C(q') + M_C\|q - q'\|_*.\end{aligned}$$

Interchanging q and q' gives

$$|h_C(q) - h_C(q')| \leq M_C\|q - q'\|_*.$$

Therefore, each support function h_{D_i} is Lipschitz continuous. By its definition in Eq. (5), F_τ is obtained from finitely many continuous support functions through exponentiation, finite summation, logarithm, and multiplication by $\tau > 0$. Since the argument of the logarithm is strictly positive, F_τ is continuous on V^*.

Let $q, q' \in V^*$ and $\theta \in [0, 1]$. For every $i \in [N]$,

$$\begin{aligned}h_{D_i}(\theta q + (1 - \theta)q') &= \max_{x\in D_i} \left(\theta\langle q, x\rangle + (1 - \theta)\langle q', x\rangle\right) \\ &\leq \theta \max_{x\in D_i}\langle q, x\rangle + (1 - \theta) \max_{x\in D_i}\langle q', x\rangle \\ &= \theta h_{D_i}(q) + (1 - \theta)h_{D_i}(q').\end{aligned}$$

Suppose first that $\theta \in (0,1)$, and define $x_i := \exp(h_{D_i}(q)/\tau)$ and $y_i := \exp(h_{D_i}(q')/\tau)$. The preceding inequality and Hölder's inequality with conjugate exponents $1/\theta$ and $1/(1-\theta)$ (Rudin, 1974) give

$$\sum_{i=1}^{N} \exp\left(\frac{h_{D_i}(\theta q + (1-\theta)q')}{\tau}\right) \le \sum_{i=1}^{N} x_i^{\theta} y_i^{1-\theta} \le \left(\sum_{i=1}^{N} x_i\right)^{\theta} \left(\sum_{i=1}^{N} y_i\right)^{1-\theta}.$$

Taking logarithms and multiplying by $\tau > 0$ yields

$$F_\tau(\theta q + (1-\theta)q') \le \theta F_\tau(q) + (1-\theta) F_\tau(q').$$

For $\theta \in \{0,1\}$, the same inequality holds with equality. Hence, F_τ is convex.

Since F_τ is continuous, the set $\mathcal{S}_\varepsilon = F_\tau^{-1}((-\infty, \beta_\varepsilon])$ is closed. If $q, q' \in \mathcal{S}_\varepsilon$ and $\theta \in [0,1]$, convexity gives

$$F_\tau(\theta q + (1-\theta)q') \le \theta F_\tau(q) + (1-\theta) F_\tau(q') \le \beta_\varepsilon.$$

Therefore, $\theta q + (1-\theta)q' \in \mathcal{S}_\varepsilon$, proving that $\mathcal{S}_\varepsilon$ is convex.

For every $i \in [N]$, the definition of the support function gives $h_{D_i}(0) = 0$. Since $N \ge 1$,

$$F_\tau(0) = \tau \log\left(\sum_{i=1}^{N} \exp(0)\right) = \tau \log N \ge 0.$$

Moreover, $\varepsilon \in (0, 1/2)$ implies $\varepsilon/(1-\varepsilon) < 1$ and hence $\beta_\varepsilon < 0$. Therefore, $F_\tau(0) > \beta_\varepsilon$, so $0 \notin \mathcal{S}_\varepsilon$.

To prove outward radiality, fix $q \in \mathcal{S}_\varepsilon$. By the sublevel-set representation,

$$\sum_{i=1}^{N} \exp\left(\frac{h_{D_i}(q)}{\tau}\right) = \exp\left(\frac{F_\tau(q)}{\tau}\right) \le \exp\left(\frac{\beta_\varepsilon}{\tau}\right) = \frac{\varepsilon}{1-\varepsilon} < 1.$$

Every term in the sum is positive, so each term is strictly smaller than one. Hence, $h_{D_i}(q) < 0$ for every $i \in [N]$. For $t \ge 1$, positive homogeneity of support functions gives

$$h_{D_i}(tq) = \max_{x \in D_i} \langle tq, x\rangle = t \max_{x \in D_i} \langle q, x\rangle = t h_{D_i}(q) \le h_{D_i}(q).$$

Consequently,

$$\begin{aligned} F_\tau(tq) &= \tau \log\left(\sum_{i=1}^{N} \exp\left(\frac{t h_{D_i}(q)}{\tau}\right)\right) \\ &\le \tau \log\left(\sum_{i=1}^{N} \exp\left(\frac{h_{D_i}(q)}{\tau}\right)\right) \\ &= F_\tau(q) \le \beta_\varepsilon. \end{aligned}$$

Thus, $tq \in \mathcal{S}_\varepsilon$ for every $t \ge 1$.

Finally, suppose that $\mathcal{S}_\varepsilon$ is nonempty. Define

$$r_\varepsilon := \inf_{q \in \mathcal{S}_\varepsilon} \|q\|_*.$$

For any fixed $q_0 \in \mathcal{S}_\varepsilon$, we have $0 \le r_\varepsilon \le \|q_0\|_* < \infty$. By the definition of the infimum, there exists a sequence $(q_n)_{n \ge 1} \subset \mathcal{S}_\varepsilon$ such that $\|q_n\|_* \to r_\varepsilon$. The sequence is bounded, so finite dimensionality of V^* implies that it has a convergent subsequence $(q_{n_j})_{j \ge 1}$. Let $q_\varepsilon^\star \in V^*$ denote its limit:

$$q_{n_j} \longrightarrow q_\varepsilon^\star.$$

Closedness of $\mathcal{S}_\varepsilon$ gives $q_\varepsilon^\star \in \mathcal{S}_\varepsilon$, and continuity of the norm gives

$$\|q_\varepsilon^\star\|_* = \lim_{j \to \infty} \|q_{n_j}\|_* = r_\varepsilon.$$

If $\mathcal{S}_\varepsilon \cap \mathcal{Q}_R \neq \varnothing$, then there exists $q \in \mathcal{S}_\varepsilon$ satisfying $\|q\|_* \leq R$, and therefore $r_\varepsilon \leq R$. It follows that $r_\varepsilon \leq R_{\mathrm{crit}}(\varepsilon)$. Conversely, $q_\varepsilon^\star \in \mathcal{S}_\varepsilon \cap \mathcal{Q}_{r_\varepsilon}$, so $R_{\mathrm{crit}}(\varepsilon) \leq r_\varepsilon$. Hence,

$$R_{\mathrm{crit}}(\varepsilon) = r_\varepsilon = \|q_\varepsilon^\star\|_*.$$

It remains to prove that $q_\varepsilon^\star$ lies on the threshold level set. Since $0 \notin \mathcal{S}_\varepsilon$, we have $q_\varepsilon^\star \neq 0$. Suppose, for contradiction, that $F_\tau(q_\varepsilon^\star) < \beta_\varepsilon$, and define $\eta := (\beta_\varepsilon - F_\tau(q_\varepsilon^\star))/2 > 0$. Continuity of $t \mapsto F_\tau(tq_\varepsilon^\star)$ at $t = 1$ implies that there exists $\rho > 0$ such that

$$|t-1| < \rho \quad \Longrightarrow \quad |F_\tau(tq_\varepsilon^\star) - F_\tau(q_\varepsilon^\star)| < \eta.$$

Choose $t \in (\max\{0, 1-\rho\}, 1)$. Then

$$\begin{aligned} F_\tau(tq_\varepsilon^\star) &< F_\tau(q_\varepsilon^\star) + \eta \\ &= \frac{F_\tau(q_\varepsilon^\star) + \beta_\varepsilon}{2} \\ &< \beta_\varepsilon. \end{aligned}$$

Thus, $tq_\varepsilon^\star \in \mathcal{S}_\varepsilon$. However,

$$\|tq_\varepsilon^\star\|_* = t\|q_\varepsilon^\star\|_* < \|q_\varepsilon^\star\|_*,$$

which contradicts the minimality of $q_\varepsilon^\star$. Therefore,

$$F_\tau(q_\varepsilon^\star) = \beta_\varepsilon.$$

This completes the proof. □

C Proof of Proposition 3.3

Proof. Fix a unit query direction $u \in V^*$ and let $\delta(u) := -\max_{i\in[N]} h_{D_i}(u)$. For every $r \geq 0$, positive homogeneity of support functions gives

$$h_{D_i}(ru) = r h_{D_i}(u).$$

Combining this identity with Eq. (6) yields

$$\underline{p}_s(ru) = \left[1 + \sum_{i=1}^{N} \exp\left(\frac{r h_{D_i}(u)}{\tau}\right)\right]^{-1}.$$

Suppose first that $\delta(u) > 0$. By its definition,

$$\max_{i\in[N]} h_{D_i}(u) = -\delta(u).$$

Hence, $h_{D_i}(u) \leq -\delta(u)$ for every $i \in [N]$. Since the maximum is taken over a finite nonempty index set, there exists $j \in [N]$ such that $h_{D_j}(u) = -\delta(u)$. It follows that

$$\exp\left(-\frac{r\delta(u)}{\tau}\right) \leq \sum_{i=1}^{N} \exp\left(\frac{r h_{D_i}(u)}{\tau}\right) \leq N \exp\left(-\frac{r\delta(u)}{\tau}\right).$$

Since $x \mapsto (1+x)^{-1}$ is decreasing on $[0, \infty)$, these inequalities imply

$$\frac{1}{1 + N\exp(-r\delta(u)/\tau)} \leq \underline{p}_s(ru) \leq \frac{1}{1 + \exp(-r\delta(u)/\tau)}.$$

Because $\delta(u) > 0$, both the lower and upper bounds converge to one as $r \to \infty$. The squeeze theorem therefore gives

$$\lim_{r\to\infty} \underline{p}_s(ru) = 1.$$

To obtain a sufficient radius for membership in $\mathcal{S}_\varepsilon$, impose that the lower bound be at least $1-\varepsilon$. Successive equivalent transformations give

$$\begin{aligned}\frac{1}{1+N\exp(-r\delta(u)/\tau)} \geq 1-\varepsilon &\iff 1 \geq (1-\varepsilon)\left(1+N\exp(-r\delta(u)/\tau)\right)\\ &\iff \varepsilon \geq N(1-\varepsilon)\exp(-r\delta(u)/\tau)\\ &\iff \exp(-r\delta(u)/\tau) \leq \frac{\varepsilon}{N(1-\varepsilon)}\\ &\iff -\frac{r\delta(u)}{\tau} \leq \log\left(\frac{\varepsilon}{N(1-\varepsilon)}\right)\\ &\iff r \geq \frac{\tau}{\delta(u)}\log\left(\frac{N(1-\varepsilon)}{\varepsilon}\right).\end{aligned}$$

Thus, the ray $\{ru : r \geq 0\}$ intersects $\mathcal{S}_\varepsilon$ whenever $\delta(u) > 0$.

Conversely, suppose that $\delta(u) \leq 0$. Then $\max_{i\in[N]} h_{D_i}(u) = -\delta(u) \geq 0$, so there exists $j \in [N]$ satisfying $h_{D_j}(u) \geq 0$. For every $r \geq 0$,

$$\sum_{i=1}^{N}\exp\left(\frac{rh_{D_i}(u)}{\tau}\right) \geq \exp\left(\frac{rh_{D_j}(u)}{\tau}\right) \geq 1.$$

Consequently,

$$\underline{p}_s(ru) = \left[1+\sum_{i=1}^{N}\exp\left(\frac{rh_{D_i}(u)}{\tau}\right)\right]^{-1} \leq \frac{1}{2} < 1-\varepsilon.$$

Therefore, $ru \notin \mathcal{S}_\varepsilon$ for every $r \geq 0$, and the ray does not intersect the robust sink region. This proves the equivalence and completes the proof. □

D Proof of Theorem 4.1

Proof. We begin with the finite-dimensional Gibbs variational identity, equivalently the convex conjugacy between log-sum-exp and negative Shannon entropy (Boyd & Vandenberghe, 2004, Example 3.25). We derive the identity below in the notation used in this proof. Let $x = (x_1, \ldots, x_N) \in \mathbb{R}^N$, and define

$$Z(x) := \sum_{i=1}^{N}\exp(x_i/\tau), \quad \sigma_i(x) := \frac{\exp(x_i/\tau)}{Z(x)}.$$

Since $\tau > 0$ and $N \geq 1$, we have $Z(x) > 0$ and $\sigma_i(x) > 0$ for every $i \in [N]$. Moreover, $\sum_{i=1}^{N}\sigma_i(x) = 1$, so $\sigma(x) := (\sigma_1(x), \ldots, \sigma_N(x)) \in \Delta_N$.

For any $\pi \in \Delta_N$, define

$$\mathrm{KL}(\pi\|\sigma(x)) := \sum_{i=1}^{N}\pi_i\log\left(\frac{\pi_i}{\sigma_i(x)}\right),$$

where a term with $\pi_i = 0$ is defined to be zero. By Gibbs' inequality, $\mathrm{KL}(\pi\|\sigma(x)) \geq 0$. Since $\log\sigma_i(x) = x_i/\tau - \log Z(x)$, we obtain

$$\begin{aligned}\tau\,\mathrm{KL}(\pi\|\sigma(x)) &= \tau\sum_{i=1}^{N}\pi_i\log\pi_i - \tau\sum_{i=1}^{N}\pi_i\log\sigma_i(x)\\ &= \tau\sum_{i=1}^{N}\pi_i\log\pi_i - \sum_{i=1}^{N}\pi_i x_i + \tau\log Z(x)\sum_{i=1}^{N}\pi_i\\ &= -\tau H(\pi) - \sum_{i=1}^{N}\pi_i x_i + \tau\log Z(x).\end{aligned}$$

Rearranging gives

$$\sum_{i=1}^{N} \pi_i x_i + \tau H(\pi) = \tau \log Z(x) - \tau \,\mathrm{KL}(\pi \| \sigma(x)) \leq \tau \log Z(x).$$

Equality holds for $\pi = \sigma(x)$. Therefore,

$$\tau \log \left(\sum_{i=1}^{N} \exp(x_i/\tau) \right) = \max_{\pi \in \Delta_N} \left\{ \sum_{i=1}^{N} \pi_i x_i + \tau H(\pi) \right\}.$$

Applying this identity with $x_i = h_{D_i}(q)$ gives, for every $q \in V^*$,

$$F_\tau(q) = \max_{\pi \in \Delta_N} \left\{ \sum_{i=1}^{N} \pi_i h_{D_i}(q) + \tau H(\pi) \right\}.$$

We next combine the support functions. Fix $\pi \in \Delta_N$. By the definitions of Minkowski addition and nonnegative scalar multiplication,

$$\begin{aligned} h_{\sum_{i=1}^N \pi_i D_i}(q) &= \max_{\substack{d_i \in D_i \\ i \in [N]}} \left\langle q, \sum_{i=1}^{N} \pi_i d_i \right\rangle \\ &= \max_{\substack{d_i \in D_i \\ i \in [N]}} \sum_{i=1}^{N} \pi_i \langle q, d_i \rangle \\ &= \sum_{i=1}^{N} \pi_i \max_{d_i \in D_i} \langle q, d_i \rangle \\ &= \sum_{i=1}^{N} \pi_i h_{D_i}(q). \end{aligned}$$

The third equality is exact because the variables $d_1, \ldots, d_N$ range independently over the product $D_1 \times \cdots \times D_N$, and compactness of each D_i ensures that every maximum is attained.

We now prove that

$$\sum_{i=1}^{N} \pi_i D_i = B_\pi - A.$$

Let $x \in \sum_{i=1}^{N} \pi_i D_i$. Then there exist $a_i \in A$ and $b_i \in B_i$ such that

$$x = \sum_{i=1}^{N} \pi_i (b_i - a_i) = \sum_{i=1}^{N} \pi_i b_i - \sum_{i=1}^{N} \pi_i a_i.$$

By definition, $\sum_{i=1}^{N} \pi_i b_i \in B_\pi$. Since A is convex, $\pi_i \geq 0$, and $\sum_{i=1}^{N} \pi_i = 1$, we also have $\sum_{i=1}^{N} \pi_i a_i \in A$. Hence, $x \in B_\pi - A$, and therefore

$$\sum_{i=1}^{N} \pi_i D_i \subseteq B_\pi - A.$$

Conversely, let $x \in B_\pi - A$. Then there exist $a \in A$ and $b_i \in B_i$ such that

$$x = \sum_{i=1}^{N} \pi_i b_i - a.$$

Using $\sum_{i=1}^{N} \pi_i = 1$, we obtain

$$x = \sum_{i=1}^{N} \pi_i b_i - \sum_{i=1}^{N} \pi_i a = \sum_{i=1}^{N} \pi_i (b_i - a).$$

Since $b_i - a \in D_i$ for every $i \in [N]$, it follows that $x \in \sum_{i=1}^N \pi_i D_i$. Thus,

$$B_\pi - A \subseteq \sum_{i=1}^N \pi_i D_i.$$

The two inclusions prove the claimed equality. Consequently,

$$\sum_{i=1}^N \pi_i h_{D_i}(q) = h_{\sum_{i=1}^N \pi_i D_i}(q) = h_{B_\pi - A}(q).$$

It follows that

$$F_\tau(q) = \max_{\pi \in \Delta_N} \left\{ h_{B_\pi - A}(q) + \tau H(\pi) \right\}.$$

Define $L_R : \mathcal{Q}_R \times \Delta_N \to \mathbb{R}$ by

$$L_R(q, \pi) := \sum_{i=1}^N \pi_i h_{D_i}(q) + \tau H(\pi).$$

For every fixed $\pi \in \Delta_N$, the function $q \mapsto L_R(q, \pi)$ is continuous and convex because it is a nonnegative weighted sum of continuous convex support functions plus a constant. For every fixed $q \in \mathcal{Q}_R$, the function $\pi \mapsto L_R(q, \pi)$ is continuous and concave because its first term is linear in π and Shannon entropy is continuous and concave on Δ_N.

The set $\mathcal{Q}_R$ is nonempty, compact, and convex because V^* is finite-dimensional and $\mathcal{Q}_R$ is a closed norm ball. The simplex Δ_N is also nonempty, compact, and convex. Hence, all conditions of Sion's minimax theorem are satisfied, and

$$\begin{aligned}
\Psi_R &= \min_{q \in \mathcal{Q}_R} F_\tau(q) \\
&= \min_{q \in \mathcal{Q}_R} \max_{\pi \in \Delta_N} L_R(q, \pi) \\
&= \max_{\pi \in \Delta_N} \min_{q \in \mathcal{Q}_R} L_R(q, \pi) \\
&= \max_{\pi \in \Delta_N} \left\{ \tau H(\pi) + \min_{q \in \mathcal{Q}_R} h_{B_\pi - A}(q) \right\}.
\end{aligned}$$

It remains to evaluate the inner minimum. Fix $\pi \in \Delta_N$ and define $K_\pi := B_\pi - A$. The set B_π is the continuous image of the compact product $B_1 \times \cdots \times B_N$ under the map $(b_1, \ldots, b_N) \mapsto \sum_{i=1}^N \pi_i b_i$, so B_π is compact. It is convex because each B_i is convex and the coefficients π_i are nonnegative. Since A is nonempty, compact, and convex, the set $K_\pi = B_\pi - A$ is also nonempty, compact, and convex.

The function $(q, x) \mapsto \langle q, x \rangle$ is continuous and bilinear on the nonempty compact convex set $\mathcal{Q}_R \times K_\pi$. For fixed x, it is convex in q, and for fixed q, it is concave in x. A second application of Sion's minimax theorem therefore gives

$$\begin{aligned}
\min_{q \in \mathcal{Q}_R} h_{K_\pi}(q) &= \min_{q \in \mathcal{Q}_R} \max_{x \in K_\pi} \langle q, x \rangle \\
&= \max_{x \in K_\pi} \min_{q \in \mathcal{Q}_R} \langle q, x \rangle.
\end{aligned}$$

By the dual representation of the norm, which follows from the Hahn–Banach theorem (Narici & Beckenstein, 1997), every $x \in V$ satisfies

$$\max_{\|q\|_* \le 1} \langle q, x \rangle = \|x\|.$$

Because the dual unit ball is centrally symmetric,

$$\min_{\|q\|_* \le 1} \langle q, x \rangle = -\|x\|.$$

Since $\mathcal{Q}_R$ is the dual unit ball scaled by R, for every $R \ge 0$,

$$\min_{q \in \mathcal{Q}_R} \langle q, x \rangle = -R\|x\|.$$

Substituting this identity into the preceding minimax formula yields

$$\begin{aligned}\min_{q\in\mathcal{Q}_R} h_{K_\pi}(q) &= \max_{x\in K_\pi}\big(-R\|x\|\big)\\ &= -R\min_{x\in K_\pi}\|x\|\\ &= -R\,\operatorname{dist}(0,K_\pi).\end{aligned}$$

By the definition of K_π,

$$\begin{aligned}\operatorname{dist}(0,K_\pi) &= \min_{x\in B_\pi - A}\|x\|\\ &= \min_{\substack{a\in A\\ b\in B_\pi}}\|b-a\|\\ &= \operatorname{dist}(A,B_\pi)\\ &= d(\pi).\end{aligned}$$

Therefore,

$$\min_{q\in\mathcal{Q}_R} h_{B_\pi - A}(q) = -R\,d(\pi).$$

Substituting this equality into the representation of Ψ_R gives

$$\Psi_R = \max_{\pi\in\Delta_N}\left\{\tau H(\pi) - R\,d(\pi)\right\}.$$

It remains to verify that the maximum over Δ_N is attained and unique. Define

$$\mathcal{U} := A\times B_1\times\cdots\times B_N$$

and, for each $\pi\in\Delta_N$, define $f_\pi:\mathcal{U}\to\mathbb{R}$ by

$$f_\pi(a,b_{1:N}) := \left\|a-\sum_{i=1}^N \pi_i b_i\right\|.$$

By the definition of B_π,

$$d(\pi) = \min_{z\in\mathcal{U}} f_\pi(z).$$

For each $i\in[N]$, define $M_i := \max_{b\in B_i}\|b\| < \infty$. For any $\pi,\pi'\in\Delta_N$ and any $z=(a,b_{1:N})\in\mathcal{U}$, the reverse triangle inequality gives

$$\begin{aligned}|f_\pi(z) - f_{\pi'}(z)| &\le \left\|\sum_{i=1}^N(\pi_i-\pi_i')b_i\right\|\\ &\le \sum_{i=1}^N|\pi_i-\pi_i'|\,\|b_i\|\\ &\le \sum_{i=1}^N M_i|\pi_i-\pi_i'|.\end{aligned}$$

Let $z_{\pi'}\in\mathcal{U}$ be a minimizer of $f_{\pi'}$, whose existence follows from compactness of $\mathcal{U}$ and continuity of $f_{\pi'}$. Then

$$\begin{aligned}d(\pi) &= \min_{z\in\mathcal{U}} f_\pi(z)\\ &\le f_\pi(z_{\pi'})\\ &\le f_{\pi'}(z_{\pi'}) + \sum_{i=1}^N M_i|\pi_i-\pi_i'|\\ &= d(\pi') + \sum_{i=1}^N M_i|\pi_i-\pi_i'|.\end{aligned}$$

Interchanging π and π' gives

$$d(\pi') \le d(\pi) + \sum_{i=1}^{N} M_i |\pi_i - \pi'_i|.$$

Combining the two inequalities yields

$$|d(\pi) - d(\pi')| \le \sum_{i=1}^{N} M_i |\pi_i - \pi'_i|.$$

Thus, d is Lipschitz continuous on Δ_N. Since H is continuous on Δ_N, the function $\pi \mapsto \tau H(\pi) - R\, d(\pi)$ is continuous. Compactness of Δ_N therefore implies that its maximum is attained.

It remains to prove uniqueness. Let $\pi, \pi' \in \Delta_N$ and $\theta \in (0,1)$. Convexity of each B_i gives

$$\begin{aligned} B_{\theta\pi+(1-\theta)\pi'} &= \sum_{i=1}^{N} \big(\theta\pi_i + (1-\theta)\pi'_i\big) B_i \\ &= \sum_{i=1}^{N} (\theta\pi_i B_i + (1-\theta)\pi'_i B_i) \\ &= \theta B_\pi + (1-\theta) B_{\pi'}. \end{aligned}$$

Choose $a \in A$ and $b \in B_\pi$ satisfying $|a-b| = d(\pi)$, and choose $a' \in A$ and $b' \in B_{\pi'}$ satisfying $|a'-b'| = d(\pi')$. Such points exist because the sets are compact. Convexity of A gives $\theta a + (1-\theta)a' \in A$, while the preceding identity gives $\theta b + (1-\theta)b' \in B_{\theta\pi+(1-\theta)\pi'}$. Therefore,

$$\begin{aligned} d\big(\theta\pi + (1-\theta)\pi'\big) &\le |\theta(a-b) + (1-\theta)(a'-b')| \\ &\le \theta|a-b| + (1-\theta)|a'-b'| \\ &= \theta d(\pi) + (1-\theta) d(\pi'). \end{aligned}$$

Hence, d is convex on Δ_N, so $\pi \mapsto -R\, d(\pi)$ is concave. Since $\tau > 0$ and H is strictly concave on Δ_N, the function $\pi \mapsto \tau H(\pi) - R\, d(\pi)$ is strictly concave. It therefore has at most one maximizer. Combined with attainment, this proves that the maximum is attained at a unique mixture $\pi^\star_R$. This completes the proof. □

E Proof of Corollary 4.3

Proof. Recall that

$$\lambda_\varepsilon := \frac{1-\varepsilon}{\varepsilon} > 1, \quad \beta_\varepsilon = -\tau \log \lambda_\varepsilon.$$

Since F_τ is continuous and $\mathcal{Q}_R$ is compact, the minimum defining Ψ_R is attained. By Proposition 3.1,

$$\begin{aligned} \mathcal{S}_\varepsilon \cap \mathcal{Q}_R \neq \varnothing &\iff \min_{q \in \mathcal{Q}_R} F_\tau(q) \le \beta_\varepsilon \\ &\iff \Psi_R \le -\tau \log \lambda_\varepsilon. \end{aligned}$$

Applying Theorem 4.1 gives

$$\begin{aligned} \mathcal{S}_\varepsilon \cap \mathcal{Q}_R \neq \varnothing &\iff \max_{\pi \in \Delta_N} \{\tau H(\pi) - R\, d(\pi)\} \le -\tau \log \lambda_\varepsilon \\ &\iff \tau H(\pi) - R\, d(\pi) \le -\tau \log \lambda_\varepsilon \quad \text{for every } \pi \in \Delta_N \\ &\iff R\, d(\pi) \ge \tau\big(H(\pi) + \log \lambda_\varepsilon\big) \quad \text{for every } \pi \in \Delta_N. \end{aligned}$$

Because $\lambda_\varepsilon > 1$ and $H(\pi) \ge 0$, we have

$$H(\pi) + \log \lambda_\varepsilon > 0$$

for every $\pi \in \Delta_N$. If $d(\pi) = 0$ for some π, the corresponding inequality becomes

$$0 \ge \tau\big(H(\pi) + \log \lambda_\varepsilon\big),$$

which is impossible. We therefore adopt the convention

$$\frac{H(\pi)+\log\lambda_\varepsilon}{d(\pi)} := +\infty \quad \text{whenever } d(\pi)=0.$$

If $d(\pi)>0$, then

$$R\,d(\pi) \geq \tau\big(H(\pi)+\log\lambda_\varepsilon\big) \iff R \geq \tau\frac{H(\pi)+\log\lambda_\varepsilon}{d(\pi)}.$$

It follows that

$$\begin{aligned}\mathcal{S}_\varepsilon \cap \mathcal{Q}_R \neq \varnothing &\iff R \geq \tau\frac{H(\pi)+\log\lambda_\varepsilon}{d(\pi)} \quad \text{for every } \pi\in\Delta_N \\ &\iff R \geq \tau \sup_{\pi\in\Delta_N}\frac{H(\pi)+\log\lambda_\varepsilon}{d(\pi)}.\end{aligned}$$

We next prove that this supremum is attained. The functions H and d are continuous on the compact simplex Δ_N. If $d(\pi_0)=0$ for some $\pi_0\in\Delta_N$, then the ratio equals $+\infty$ at π_0 by the stated convention, so the supremum is attained. Suppose instead that $d(\pi)>0$ for every $\pi\in\Delta_N$. Continuity of d and compactness of Δ_N then give

$$\min_{\pi\in\Delta_N} d(\pi) > 0.$$

Hence, the function

$$\pi \longmapsto \frac{H(\pi)+\log\lambda_\varepsilon}{d(\pi)}$$

is continuous on Δ_N and therefore attains its maximum. Thus, in both cases,

$$\sup_{\pi\in\Delta_N}\frac{H(\pi)+\log\lambda_\varepsilon}{d(\pi)} = \max_{\pi\in\Delta_N}\frac{H(\pi)+\log\lambda_\varepsilon}{d(\pi)}.$$

If the maximum is finite, the feasible query-norm bounds form the interval beginning at its value; if the maximum is infinite, no finite query-norm bound is sufficient. Taking the infimum over all sufficient query-norm bounds and using the convention $\inf\varnothing=+\infty$ yields

$$R_{\mathrm{crit}}(\varepsilon) = \tau\max_{\pi\in\Delta_N}\frac{H(\pi)+\log\lambda_\varepsilon}{d(\pi)}.$$

This completes the proof. □

F Proof of Corollary 4.5

Proof. Recall that

$$B := \mathrm{conv}\left(\bigcup_{i=1}^N B_i\right), \quad \gamma := \mathrm{dist}(A,B).$$

Because N is finite and each B_i is compact, the union $\bigcup_{i=1}^N B_i$ is compact. Since V is finite-dimensional, the convex hull of a compact set is compact. Hence, B is nonempty, compact, and convex.

We first prove that B is the union of all competitor mixture bodies:

$$B = \bigcup_{\pi\in\Delta_N} B_\pi.$$

Fix $\pi\in\Delta_N$ and $b\in B_\pi$. By the definition of B_π, there exist $b_i\in B_i$ such that

$$b = \sum_{i=1}^N \pi_i b_i.$$

This is a convex combination of points in $\bigcup_{i=1}^{N} B_i$, so $b \in B$. Therefore, $B_\pi \subseteq B$ for every $\pi \in \Delta_N$, and

$$\bigcup_{\pi \in \Delta_N} B_\pi \subseteq B.$$

Conversely, fix $b \in B$. By the definition of the convex hull, there exist an integer $m \geq 1$, coefficients $\alpha_1, \ldots, \alpha_m \geq 0$ satisfying $\sum_{\ell=1}^{m} \alpha_\ell = 1$, indices $i_1, \ldots, i_m \in [N]$, and points $x_\ell \in B_{i_\ell}$ such that

$$b = \sum_{\ell=1}^{m} \alpha_\ell x_\ell.$$

For each $i \in [N]$, define

$$\pi_i := \sum_{\ell : i_\ell = i} \alpha_\ell.$$

Then $\pi_i \geq 0$ and

$$\sum_{i=1}^{N} \pi_i = \sum_{\ell=1}^{m} \alpha_\ell = 1,$$

so $\pi := (\pi_1, \ldots, \pi_N) \in \Delta_N$.

If $\pi_i > 0$, define

$$b_i := \frac{1}{\pi_i} \sum_{\ell : i_\ell = i} \alpha_\ell x_\ell.$$

The coefficients α_ℓ / π_i in this sum are nonnegative and sum to one over the indices satisfying $i_\ell = i$. Since every corresponding x_ℓ belongs to B_i and B_i is convex, we have $b_i \in B_i$. If $\pi_i = 0$, choose any $b_i \in B_i$, which is possible because B_i is nonempty. Then

$$\begin{aligned}
\sum_{i=1}^{N} \pi_i b_i &= \sum_{i : \pi_i > 0} \pi_i \left(\frac{1}{\pi_i} \sum_{\ell : i_\ell = i} \alpha_\ell x_\ell \right) \\
&= \sum_{i : \pi_i > 0} \sum_{\ell : i_\ell = i} \alpha_\ell x_\ell \\
&= \sum_{\ell=1}^{m} \alpha_\ell x_\ell \\
&= b.
\end{aligned}$$

Thus, $b \in B_\pi$, proving

$$B \subseteq \bigcup_{\pi \in \Delta_N} B_\pi.$$

The two inclusions establish the claimed equality.

Since $B_\pi \subseteq B$ for every $\pi \in \Delta_N$,

$$d(\pi) = \operatorname{dist}(A, B_\pi) \geq \operatorname{dist}(A, B) = \gamma.$$

Compactness of A and B implies that the distance between them is attained. Hence, there exist $a^\star \in A$ and $b^\star \in B$ such that

$$\|a^\star - b^\star\| = \gamma.$$

Since $B = \bigcup_{\pi \in \Delta_N} B_\pi$, there exists $\pi^\star \in \Delta_N$ such that $b^\star \in B_{\pi^\star}$. Therefore,

$$d(\pi^\star) \leq \|a^\star - b^\star\| = \gamma.$$

Together with $d(\pi^\star) \geq \gamma$, this gives

$$d(\pi^\star) = \gamma.$$

For every $\pi \in \Delta_N$, Shannon entropy satisfies

$$0 \leq H(\pi) \leq \log N.$$

The lower bound follows because $-\pi_i \log \pi_i \geq 0$ for every $\pi_i \in [0,1]$, with $0 \log 0 := 0$. To prove the upper bound, define the uniform vector $\nu := (1/N, \ldots, 1/N) \in \Delta_N$. Nonnegativity of the Kullback–Leibler divergence gives

$$\begin{aligned} 0 \leq \mathrm{KL}(\pi \| \nu) &= \sum_{i=1}^{N} \pi_i \log \left(\frac{\pi_i}{1/N} \right) \\ &= \sum_{i=1}^{N} \pi_i \log \pi_i + \log N \sum_{i=1}^{N} \pi_i \\ &= \log N - H(\pi). \end{aligned}$$

Hence, $H(\pi) \leq \log N$.

Using $d(\pi) \geq \gamma$ and $H(\pi) \leq \log N$ in Theorem 4.1, we obtain

$$\begin{aligned} \Psi_R &= \max_{\pi \in \Delta_N} \{\tau H(\pi) - R\, d(\pi)\} \\ &\leq \max_{\pi \in \Delta_N} \{\tau \log N - R\gamma\} \\ &= \tau \log N - R\gamma. \end{aligned}$$

Evaluating the same maximum at $\pi^\star$ and using $H(\pi^\star) \geq 0$ gives

$$\begin{aligned} \Psi_R &\geq \tau H(\pi^\star) - R\, d(\pi^\star) \\ &= \tau H(\pi^\star) - R\gamma \\ &\geq -R\gamma. \end{aligned}$$

Therefore,

$$-R\gamma \leq \Psi_R \leq \tau \log N - R\gamma.$$

The function $g_\tau : \mathbb{R} \to (0,1)$ defined by $g_\tau(x) := (1+\exp(x/\tau))^{-1}$ is strictly decreasing. Applying g_τ to the preceding inequalities and using Eq. (8) yields

$$\frac{1}{1 + N \exp(-R\gamma/\tau)} \leq P_R^* \leq \frac{1}{1 + \exp(-R\gamma/\tau)}.$$

Suppose first that $\gamma > 0$. Then $\exp(-R\gamma/\tau) \to 0$ as $R \to \infty$, so both bounds converge to one. The squeeze theorem gives

$$\lim_{R \to \infty} P_R^* = 1.$$

If $\gamma = 0$, the upper bound becomes

$$P_R^* \leq \frac{1}{2}$$

for every $R \geq 0$, so P_R^* cannot converge to one. Therefore,

$$\lim_{R \to \infty} P_R^* = 1 \iff \gamma > 0.$$

Since the distance between the compact sets A and B is attained, $\gamma = 0$ implies that there exist $a \in A$ and $b \in B$ satisfying $\|a - b\| = 0$, hence $a = b$ and $A \cap B \neq \varnothing$. Conversely, if $A \cap B \neq \varnothing$, then $\gamma = 0$. Thus,

$$\gamma > 0 \iff A \cap B = \varnothing.$$

It remains to compute the asymptotic error exponent. Define

$$x_R := \exp(-R\gamma/\tau).$$

Since $\gamma \geq 0$, we have $0 < x_R \leq 1$. Subtracting the probability bounds from one gives

$$\frac{x_R}{1 + x_R} \leq 1 - P_R^* \leq \frac{N x_R}{1 + N x_R}.$$

The logarithm is increasing, so multiplying the logarithmic inequalities by $-\tau/R < 0$ reverses their directions:

$$-\frac{\tau}{R}\log\left(\frac{Nx_R}{1+Nx_R}\right) \le -\frac{\tau}{R}\log(1-P_R^*) \le -\frac{\tau}{R}\log\left(\frac{x_R}{1+x_R}\right).$$

Using $\log x_R = -R\gamma/\tau$, the two outer terms can be written as

$$-\frac{\tau}{R}\log\left(\frac{Nx_R}{1+Nx_R}\right) = \gamma - \frac{\tau\log N}{R} + \frac{\tau}{R}\log(1+Nx_R),$$
$$-\frac{\tau}{R}\log\left(\frac{x_R}{1+x_R}\right) = \gamma + \frac{\tau}{R}\log(1+x_R).$$

Since $0 < x_R \le 1$,

$$0 \le \log(1+x_R) \le \log 2, \quad 0 \le \log(1+Nx_R) \le \log(1+N).$$

Therefore,

$$\lim_{R\to\infty}\frac{\tau}{R}\log(1+x_R) = 0, \quad \lim_{R\to\infty}\frac{\tau}{R}\log(1+Nx_R) = 0, \quad \lim_{R\to\infty}\frac{\tau\log N}{R} = 0.$$

Both outer terms converge to γ. The squeeze theorem consequently yields

$$\lim_{R\to\infty} -\frac{\tau}{R}\log(1-P_R^*) = \gamma.$$

This completes the proof. □

G Proof of Theorem 5.1

Proof. We prove the three claims of Theorem 5.1 in four steps.

Step 1: Reduction of the full competitor simplex. Fix an integer $m \ge 1$ and a perturbation radius $\rho \ge 0$. Since V is finite-dimensional, its closed unit ball $\mathbb{B}$ is compact. Therefore, the uncertainty sets A_ρ, $B_{1,\rho}, \dots, B_{n,\rho}$, and C_ρ defined in Assumption 2.9 are nonempty, compact, and convex. Hence, Corollary 4.3 applies to the resulting attention problem with $n+m$ competitors.

A mixture over the n existing competitors and the m added competitors is represented by

$$\omega := (\pi_1, \dots, \pi_n, \eta_1, \dots, \eta_m) \in \Delta_{n+m}.$$

Here, π_i is the mixture weight of existing competitor $i \in [n]$, and η_j is the mixture weight of added competitor $j \in [m]$. Define the total weight assigned to the added competitors by

$$\alpha := \sum_{j=1}^{m} \eta_j.$$

Since $\omega \in \Delta_{n+m}$,

$$\sum_{i=1}^{n} \pi_i + \alpha = 1.$$

Consequently, $z := (\pi, \alpha)$, where $\pi := (\pi_1, \dots, \pi_n)$, belongs to the reduced simplex $\mathcal{Z}$.

The competitor mixture set associated with ω is

$$\widetilde{B}_\omega := \sum_{i=1}^{n} \pi_i(k_i + \rho\mathbb{B}) + \sum_{j=1}^{m} \eta_j(C + \rho\mathbb{B}).$$

For a set $S \subset V$ and a scalar $t \ge 0$, write $tS := \{ts : s \in S\}$, with $0S = \{0\}$. We first record that, for every nonempty convex set $S \subset V$ and nonnegative scalars $t_1, \dots, t_\ell$,

$$\sum_{j=1}^{\ell} t_j S = \left(\sum_{j=1}^{\ell} t_j\right) S.$$

To verify this identity, let $t := \sum_{j=1}^{\ell} t_j$. If $t = 0$, then $t_j = 0$ for every j, and both sides equal $\{0\}$. Suppose that $t > 0$. For arbitrary $s_1, \dots, s_\ell \in S$, convexity of S gives

$$\sum_{j=1}^{\ell} t_j s_j = t \sum_{j=1}^{\ell} \frac{t_j}{t} s_j \in tS.$$

This proves $\sum_{j=1}^{\ell} t_j S \subseteq tS$. Conversely, every $ts \in tS$ satisfies

$$ts = \sum_{j=1}^{\ell} t_j s \in \sum_{j=1}^{\ell} t_j S,$$

which proves the reverse inclusion.

Applying this identity to C and $\mathbb{B}$ gives

$$\begin{aligned}
\widetilde{B}_\omega &= \sum_{i=1}^{n} \pi_i k_i + \rho \sum_{i=1}^{n} \pi_i \mathbb{B} + \sum_{j=1}^{m} \eta_j C + \rho \sum_{j=1}^{m} \eta_j \mathbb{B} \\
&= \sum_{i=1}^{n} \pi_i k_i + \rho \left(\sum_{i=1}^{n} \pi_i \right) \mathbb{B} + \alpha C + \rho\alpha \mathbb{B} \\
&= \sum_{i=1}^{n} \pi_i k_i + \alpha C + \rho \left(\sum_{i=1}^{n} \pi_i + \alpha \right) \mathbb{B} \\
&= K_z + \rho \mathbb{B}.
\end{aligned}$$

We next establish the distance identity

$$\operatorname{dist}(k_s + \rho\mathbb{B}, K + \rho\mathbb{B}) = (\operatorname{dist}(\{k_s\}, K) - 2\rho)_+$$

for every nonempty compact set $K \subset V$. Define

$$d := \operatorname{dist}(\{k_s\}, K) = \min_{k \in K} \|k_s - k\|.$$

For arbitrary $u, v \in \mathbb{B}$ and $k \in K$, the reverse triangle inequality and the triangle inequality give

$$\begin{aligned}
\|k_s + \rho u - (k + \rho v)\| &= \|(k_s - k) + \rho(u - v)\| \\
&\geq \|k_s - k\| - \rho\|u - v\| \\
&\geq \|k_s - k\| - \rho\|u\| - \rho\|v\| \\
&\geq d - 2\rho.
\end{aligned}$$

Since norms are nonnegative,

$$\|k_s + \rho u - (k + \rho v)\| \geq (d - 2\rho)_+.$$

Taking the infimum over $u, v \in \mathbb{B}$ and $k \in K$ yields

$$\operatorname{dist}(k_s + \rho\mathbb{B}, K + \rho\mathbb{B}) \geq (d - 2\rho)_+.$$

Compactness of K provides $k^\star \in K$ satisfying $\|k^\star - k_s\| = d$. If $d = 0$, then $k^\star = k_s$, so k_s belongs to both $k_s + \rho\mathbb{B}$ and $K + \rho\mathbb{B}$. Hence, the distance between these sets is zero, which equals $(d - 2\rho)_+$.

Suppose that $d > 0$, and define

$$e := \frac{k^\star - k_s}{d}.$$

Then $\|e\| = 1$. If $d > 2\rho$, define

$$a := k_s + \rho e, \quad b := k^\star - \rho e.$$

Since $e, -e \in \mathbb{B}$, we have $a \in k_s + \rho\mathbb{B}$ and $b \in K + \rho\mathbb{B}$. Moreover,

$$\begin{aligned}
b - a &= k^\star - k_s - 2\rho e \\
&= (d - 2\rho)e,
\end{aligned}$$

and therefore

$$\|b - a\| = d - 2\rho.$$

Thus,

$$\operatorname{dist}(k_s + \rho\mathbb{B}, K + \rho\mathbb{B}) \le d - 2\rho = (d - 2\rho)_+.$$

If $0 < d \le 2\rho$, then $\rho > 0$. Define

$$a := k_s + \frac{d}{2}e, \quad b := k^\star - \frac{d}{2}e.$$

Since $d/(2\rho) \le 1$, we have

$$a = k_s + \rho\left(\frac{d}{2\rho}e\right) \in k_s + \rho\mathbb{B}, \quad b = k^\star + \rho\left(-\frac{d}{2\rho}e\right) \in K + \rho\mathbb{B}.$$

Furthermore,

$$a = k_s + \frac{k^\star - k_s}{2} = \frac{k_s + k^\star}{2} = k^\star - \frac{k^\star - k_s}{2} = b.$$

Hence, the two expanded sets intersect, and their distance is zero. Since $(d - 2\rho)_+ = 0$, the reverse inequality also holds in this case. We have therefore proved

$$\operatorname{dist}(k_s + \rho\mathbb{B}, K + \rho\mathbb{B}) = \left(\operatorname{dist}(\{k_s\}, K) - 2\rho\right)_+.$$

Applying this identity with $K = K_z$ yields

$$\operatorname{dist}(A_\rho, \widetilde{B}_\omega) = \operatorname{dist}(k_s + \rho\mathbb{B}, K_z + \rho\mathbb{B}) = (r(z) - 2\rho)_+.$$

Thus, the distance depends on $\eta_1, \dots, \eta_m$ only through their sum α.

The entropy of the full mixture ω is

$$H(\omega) = -\sum_{i=1}^{n} \pi_i \log \pi_i - \sum_{j=1}^{m} \eta_j \log \eta_j.$$

If $\alpha = 0$, then $\eta_j = 0$ for every $j \in [m]$, and

$$-\sum_{j=1}^{m} \eta_j \log \eta_j = 0 = -\alpha \log \alpha + \alpha \log m.$$

Suppose that $\alpha > 0$, and define $\xi_j := \eta_j/\alpha$ for every $j \in [m]$. Then $\xi := (\xi_1, \dots, \xi_m) \in \Delta_m$, and

$$\begin{aligned} -\sum_{j=1}^{m} \eta_j \log \eta_j &= -\sum_{j=1}^{m} \alpha\xi_j \log(\alpha\xi_j) \\ &= -\alpha \log \alpha \sum_{j=1}^{m} \xi_j - \alpha \sum_{j=1}^{m} \xi_j \log \xi_j \\ &= -\alpha \log \alpha + \alpha H(\xi). \end{aligned}$$

Let $\upsilon_m := (1/m, \dots, 1/m) \in \Delta_m$. Nonnegativity of the Kullback–Leibler divergence gives

$$\begin{aligned} 0 \le \mathrm{KL}(\xi \| \upsilon_m) &= \sum_{j=1}^{m} \xi_j \log\left(\frac{\xi_j}{1/m}\right) \\ &= \sum_{j=1}^{m} \xi_j \log \xi_j + \log m \sum_{j=1}^{m} \xi_j \\ &= \log m - H(\xi). \end{aligned}$$

Hence, $H(\xi) \le \log m$, with equality if and only if $\xi = \upsilon_m$. Consequently,

$$-\sum_{j=1}^{m} \eta_j \log \eta_j \le -\alpha \log \alpha + \alpha \log m,$$

with equality at $\eta_j = \alpha/m$ for every $j \in [m]$. Therefore, for every fixed $z = (\pi, \alpha) \in \mathcal{Z}$,

$$\max_{\substack{\eta_j \geq 0 \\ \sum_{j=1}^m \eta_j = \alpha}} H(\omega) = H_0(z) + \alpha \log m.$$

Applying Corollary 4.3 to the $n + m$ competitor bodies gives

$$R_{\mathrm{crit}}(m, \rho; \varepsilon) = \tau \sup_{\omega \in \Delta_{n+m}} \frac{H(\omega) + \log \lambda_\varepsilon}{\mathrm{dist}(A_\rho, \widetilde{B}_\omega)}.$$

For $z \in \mathcal{Z}$, define

$$D_\rho(z) := (r(z) - 2\rho)_+.$$

If $D_\rho(z) = 0$, then $H(\omega) + \log \lambda_\varepsilon > 0$ for every full mixture ω inducing z, because $H(\omega) \geq 0$ and $\log \lambda_\varepsilon > 0$. By the convention in Corollary 4.3, every corresponding ratio equals $+\infty$. If $D_\rho(z) > 0$, maximizing the ratio over all $\eta_1, \ldots, \eta_m$ with total mass α is equivalent to maximizing $H(\omega)$, which is achieved by $\eta_j = \alpha/m$. Therefore, in both cases,

$$R_{\mathrm{crit}}(m, \rho; \varepsilon) = \tau \sup_{z=(\pi,\alpha) \in \mathcal{Z}} \frac{H_0(z) + \alpha \log m + \log \lambda_\varepsilon}{(r(z) - 2\rho)_+}.$$

This proves part *(i)*.

Step 2: Identification of the phase boundary. Define the aggregate nominal competitor set by

$$K := \mathrm{conv}\left(\{k_1, \ldots, k_n\} \cup C\right).$$

We first prove

$$K = \bigcup_{z \in \mathcal{Z}} K_z.$$

Let $z = (\pi, \alpha) \in \mathcal{Z}$ and $y \in K_z$. By the definition of K_z, there exists $c \in C$ such that

$$y = \sum_{i=1}^n \pi_i k_i + \alpha c.$$

Since $\pi_i \geq 0$, $\alpha \geq 0$, and $\sum_{i=1}^n \pi_i + \alpha = 1$, this is a convex combination of points in $\{k_1, \ldots, k_n\} \cup C$. Hence, $y \in K$, proving

$$\bigcup_{z \in \mathcal{Z}} K_z \subseteq K.$$

Conversely, let $y \in K$. By the definition of the convex hull, y is a finite convex combination of points in $\{k_1, \ldots, k_n\} \cup C$. After aggregating all coefficients assigned to the same point k_i, there exist coefficients $\pi_1, \ldots, \pi_n \geq 0$, an integer $\ell \geq 0$, coefficients $\beta_1, \ldots, \beta_\ell \geq 0$, and points $c_1, \ldots, c_\ell \in C$ such that

$$y = \sum_{i=1}^n \pi_i k_i + \sum_{j=1}^\ell \beta_j c_j, \quad \sum_{i=1}^n \pi_i + \sum_{j=1}^\ell \beta_j = 1.$$

Define

$$\alpha := \sum_{j=1}^\ell \beta_j.$$

If $\alpha > 0$, define

$$c := \frac{1}{\alpha} \sum_{j=1}^\ell \beta_j c_j.$$

The coefficients β_j/α are nonnegative and sum to one, so convexity of C gives $c \in C$. Therefore,

$$y = \sum_{i=1}^n \pi_i k_i + \alpha c \in K_{(\pi,\alpha)}.$$

If $\alpha = 0$, choose any $c \in C$, which is possible because C is nonempty. Then

$$y = \sum_{i=1}^{n} \pi_i k_i + 0c \in K_{(\pi,0)}.$$

Thus,

$$K \subseteq \bigcup_{z \in \mathcal{Z}} K_z,$$

and the two inclusions prove the claimed equality.

We next establish compactness and continuity before invoking any minimum. The reduced simplex $\mathcal{Z}$ is compact because it is a closed and bounded subset of the finite-dimensional space $\mathbb{R}^{n+1}$. Define $T : \mathcal{Z} \times C \to V$ by

$$T((\pi, \alpha), c) := \sum_{i=1}^{n} \pi_i k_i + \alpha c.$$

The map T is continuous, and

$$T(\mathcal{Z} \times C) = \bigcup_{z \in \mathcal{Z}} K_z = K.$$

Since $\mathcal{Z} \times C$ is compact, its continuous image K is compact.

We now prove continuity of r. Define

$$M_C := \max_{c \in C} \|c\| < \infty.$$

For $z = (\pi, \alpha) \in \mathcal{Z}$, define $f_z : C \to \mathbb{R}$ by

$$f_z(c) := \left\| k_s - \sum_{i=1}^{n} \pi_i k_i - \alpha c \right\|.$$

Then

$$r(z) = \min_{c \in C} f_z(c).$$

For $z = (\pi, \alpha)$ and $z' = (\pi', \alpha')$ in $\mathcal{Z}$, the reverse triangle inequality gives, for every $c \in C$,

$$\begin{aligned} |f_z(c) - f_{z'}(c)| &\le \left\| \sum_{i=1}^{n} (\pi_i - \pi_i') k_i + (\alpha - \alpha') c \right\| \\ &\le \sum_{i=1}^{n} |\pi_i - \pi_i'| \, \|k_i\| + |\alpha - \alpha'| \, \|c\| \\ &\le \sum_{i=1}^{n} |\pi_i - \pi_i'| \, \|k_i\| + |\alpha - \alpha'| M_C. \end{aligned}$$

Let $c_{z'} \in C$ be a minimizer of $f_{z'}$, whose existence follows from compactness of C and continuity of $f_{z'}$. Then

$$\begin{aligned} r(z) &= \min_{c \in C} f_z(c) \\ &\le f_z(c_{z'}) \\ &\le f_{z'}(c_{z'}) + \sum_{i=1}^{n} |\pi_i - \pi_i'| \, \|k_i\| + |\alpha - \alpha'| M_C \\ &= r(z') + \sum_{i=1}^{n} |\pi_i - \pi_i'| \, \|k_i\| + |\alpha - \alpha'| M_C. \end{aligned}$$

Interchanging z and z' gives

$$r(z') \le r(z) + \sum_{i=1}^{n} |\pi_i - \pi_i'| \, \|k_i\| + |\alpha - \alpha'| M_C.$$

Combining the two inequalities yields

$$|r(z) - r(z')| \le \sum_{i=1}^{n} |\pi_i - \pi_i'| \, \|k_i\| + |\alpha - \alpha'| M_C.$$

Thus, r is continuous on $\mathcal{Z}$.

Since r is continuous and $\mathcal{Z}$ is compact, it attains its minimum. Using the identity $K = \bigcup_{z \in \mathcal{Z}} K_z$, we obtain

$$\begin{aligned}
\gamma_0 &= \min_{z \in \mathcal{Z}} r(z) \\
&= \min_{z \in \mathcal{Z}} \min_{y \in K_z} \|k_s - y\| \\
&= \min_{y \in \bigcup_{z \in \mathcal{Z}} K_z} \|k_s - y\| \\
&= \min_{y \in K} \|k_s - y\| \\
&= \operatorname{dist}\left(\{k_s\}, \operatorname{conv}\left(\{k_1, \ldots, k_n\} \cup C\right)\right).
\end{aligned}$$

Moreover,

$$\mathcal{M}_{\gamma_0} = \{z \in \mathcal{Z} : r(z) = \gamma_0\}$$

is nonempty because the minimum is attained, and it is compact because it is a closed subset of the compact set $\mathcal{Z}$.

For fixed $m \ge 1$, define

$$\phi_m(z) := H_0(z) + \alpha \log m + \log \lambda_\varepsilon.$$

The function $x \mapsto -x \log x$, extended by the convention $0 \log 0 := 0$, is continuous on $[0, 1]$. Hence, H_0 and ϕ_m are continuous on $\mathcal{Z}$. Furthermore,

$$H_0(z) \ge 0, \quad \alpha \log m \ge 0, \quad \log \lambda_\varepsilon > 0,$$

so

$$\phi_m(z) > 0$$

for every $z \in \mathcal{Z}$.

If $2\rho \ge \gamma_0$, choose $z_{\gamma_0} \in \mathcal{M}_{\gamma_0}$. Then

$$(r(z_{\gamma_0}) - 2\rho)_+ = (\gamma_0 - 2\rho)_+ = 0.$$

Since $\phi_m(z_{\gamma_0}) > 0$, the convention in Eq. (12) gives

$$R_{\mathrm{crit}}(m, \rho; \varepsilon) = +\infty.$$

Suppose instead that $2\rho < \gamma_0$. Then, for every $z \in \mathcal{Z}$,

$$r(z) - 2\rho \ge \gamma_0 - 2\rho > 0.$$

To bound the numerator, let

$$\bar{v}_{n+1} := \left(\frac{1}{n+1}, \ldots, \frac{1}{n+1}\right) \in \Delta_{n+1}.$$

Since $(\pi_1, \ldots, \pi_n, \alpha) \in \Delta_{n+1}$, nonnegativity of the Kullback–Leibler divergence gives

$$\begin{aligned}
0 &\le \mathrm{KL}\left((\pi, \alpha) \| \bar{v}_{n+1}\right) \\
&= \sum_{i=1}^{n} \pi_i \log \pi_i + \alpha \log \alpha + \log(n+1) \\
&= \log(n+1) - H_0(z).
\end{aligned}$$

Therefore,

$$0 \le H_0(z) \le \log(n+1).$$

Since $0 \le \alpha \le 1$ and $\log m \ge 0$,

$$0 < \phi_m(z) \le \log(n+1) + \log m + \log \lambda_\varepsilon.$$

Consequently,

$$\frac{\phi_m(z)}{r(z)-2\rho} \le \frac{\log(n+1)+\log m+\log \lambda_\varepsilon}{\gamma_0-2\rho} < \infty$$

uniformly over $z \in \mathcal{Z}$. Eq. (12) therefore implies

$$R_{\mathrm{crit}}(m,\rho;\varepsilon) < \infty.$$

We conclude that

$$R_{\mathrm{crit}}(m,\rho;\varepsilon) < \infty \quad \Longleftrightarrow \quad 2\rho < \gamma_0,$$

which proves part *(ii)*.

Step 3: Context-length asymptotics. Fix $\rho \in [0, \gamma_0/2)$ and define

$$g_\rho(z) := r(z) - 2\rho.$$

For every $z \in \mathcal{Z}$,

$$g_\rho(z) \ge \gamma_0 - 2\rho > 0.$$

The function $z \mapsto \alpha/g_\rho(z)$ is continuous on the compact set $\mathcal{Z}$, so it attains its maximum. Define

$$\kappa_\rho := \max_{z=(\pi,\alpha)\in\mathcal{Z}} \frac{\alpha}{g_\rho(z)}.$$

For every integer $m \ge 2$, Eq. (12) gives

$$\begin{aligned}\frac{R_{\mathrm{crit}}(m,\rho;\varepsilon)}{\tau \log m} &= \max_{z=(\pi,\alpha)\in\mathcal{Z}} \frac{H_0(z)+\alpha\log m+\log\lambda_\varepsilon}{g_\rho(z)\log m} \\ &= \max_{z=(\pi,\alpha)\in\mathcal{Z}} \left\{ \frac{\alpha}{g_\rho(z)} + \frac{H_0(z)+\log\lambda_\varepsilon}{g_\rho(z)\log m} \right\}.\end{aligned}$$

The second term inside the maximum is nonnegative. Hence,

$$\frac{R_{\mathrm{crit}}(m,\rho;\varepsilon)}{\tau \log m} \ge \max_{z=(\pi,\alpha)\in\mathcal{Z}} \frac{\alpha}{g_\rho(z)} = \kappa_\rho.$$

Using $H_0(z) \le \log(n+1)$ and $g_\rho(z) \ge \gamma_0 - 2\rho$, we also have

$$0 \le \frac{H_0(z)+\log\lambda_\varepsilon}{g_\rho(z)\log m} \le \frac{\log(n+1)+\log\lambda_\varepsilon}{(\gamma_0-2\rho)\log m}.$$

Therefore,

$$\frac{R_{\mathrm{crit}}(m,\rho;\varepsilon)}{\tau \log m} \le \kappa_\rho + \frac{\log(n+1)+\log\lambda_\varepsilon}{(\gamma_0-2\rho)\log m}.$$

Combining the lower and upper bounds gives

$$\kappa_\rho \le \frac{R_{\mathrm{crit}}(m,\rho;\varepsilon)}{\tau \log m} \le \kappa_\rho + \frac{\log(n+1)+\log\lambda_\varepsilon}{(\gamma_0-2\rho)\log m}.$$

The final term converges to zero as $m \to \infty$. By the squeeze theorem,

$$\lim_{\substack{m\to\infty \\ m\in\mathbb{N}}} \frac{R_{\mathrm{crit}}(m,\rho;\varepsilon)}{\tau \log m} = \kappa_\rho = \max_{z=(\pi,\alpha)\in\mathcal{Z}} \frac{\alpha}{r(z)-2\rho}.$$

This proves Eq. (14).

Step 4: Perturbation-boundary asymptotics. Assume $\gamma_0 > 0$ and fix an integer $m \ge 1$. Define

$$\phi_m(z) := H_0(z) + \alpha \log m + \log \lambda_\varepsilon, \quad g(z) := r(z) - \gamma_0.$$

Both ϕ_m and g are continuous on $\mathcal{Z}$. Moreover,

$$\phi_m(z) > 0, \quad g(z) \ge 0, \quad \mathcal{M}_{\gamma_0} = \{z \in \mathcal{Z} : g(z) = 0\}.$$

For $\rho \in [0, \gamma_0/2)$, define

$$\delta_\rho := \gamma_0 - 2\rho > 0.$$

Since $r(z) - 2\rho = \delta_\rho + g(z)$, Eq. (12) becomes

$$\frac{R_{\mathrm{crit}}(m, \rho; \varepsilon)}{\tau} = \max_{z \in \mathcal{Z}} \frac{\phi_m(z)}{\delta_\rho + g(z)}.$$

The supremum is a maximum because the objective is continuous and its denominator is strictly positive on the compact set $\mathcal{Z}$.

Define

$$c_m(\varepsilon) := \max_{z \in \mathcal{M}_{\gamma_0}} \phi_m(z).$$

The maximum is attained because $\mathcal{M}_{\gamma_0}$ is nonempty and compact. Since $\phi_m(z) \geq \log \lambda_\varepsilon > 0$, we have

$$c_m(\varepsilon) > 0.$$

Choose $z^\star \in \mathcal{M}_{\gamma_0}$ satisfying $\phi_m(z^\star) = c_m(\varepsilon)$. Since $g(z^\star) = 0$,

$$\begin{aligned} \frac{\delta_\rho R_{\mathrm{crit}}(m, \rho; \varepsilon)}{\tau} &= \max_{z \in \mathcal{Z}} \frac{\delta_\rho \phi_m(z)}{\delta_\rho + g(z)} \\ &\geq \frac{\delta_\rho \phi_m(z^\star)}{\delta_\rho + g(z^\star)} \\ &= c_m(\varepsilon). \end{aligned}$$

Therefore,

$$\liminf_{\rho \uparrow \gamma_0/2} \frac{\delta_\rho R_{\mathrm{crit}}(m, \rho; \varepsilon)}{\tau} \geq c_m(\varepsilon).$$

We now prove the reverse inequality. Let $(\rho_j)_{j \geq 1}$ be an arbitrary sequence in $[0, \gamma_0/2)$ satisfying $\rho_j \to \gamma_0/2$, and define

$$\delta_j := \gamma_0 - 2\rho_j, \quad L_j := \frac{\delta_j R_{\mathrm{crit}}(m, \rho_j; \varepsilon)}{\tau}.$$

Since

$$L_j = \max_{z \in \mathcal{Z}} \frac{\delta_j \phi_m(z)}{\delta_j + g(z)}$$

and $0 < \delta_j/(\delta_j + g(z)) \leq 1$, we have

$$0 \leq L_j \leq \max_{z \in \mathcal{Z}} \phi_m(z) < \infty.$$

Thus, $(L_j)_{j \geq 1}$ is bounded. Choose a subsequence, still indexed by j, such that

$$L_j \longrightarrow \limsup_{\ell \to \infty} L_\ell.$$

For each j, compactness of $\mathcal{Z}$ provides a maximizer $z_j \in \mathcal{Z}$ satisfying

$$L_j = \frac{\delta_j \phi_m(z_j)}{\delta_j + g(z_j)}.$$

By compactness of $\mathcal{Z}$, after passing to a further subsequence and retaining the same notation, there exists $\bar{z} \in \mathcal{Z}$ such that

$$z_j \longrightarrow \bar{z}.$$

Suppose first that $g(\bar{z}) > 0$. Continuity of g implies that there exists $j_0 \geq 1$ such that

$$g(z_j) \geq \frac{g(\bar{z})}{2}$$

for every $j \geq j_0$. Let $M_\phi := \max_{z \in \mathcal{Z}} \phi_m(z) < \infty$. Then, for every $j \geq j_0$,

$$0 \leq L_j = \frac{\delta_j \phi_m(z_j)}{\delta_j + g(z_j)} \leq \frac{\delta_j M_\phi}{g(z_j)} \leq \frac{2\delta_j M_\phi}{g(\bar{z})}.$$

Since $\delta_j \to 0$, it follows that

$$L_j \longrightarrow 0 \leq c_m(\varepsilon).$$

Suppose instead that $g(\bar{z}) = 0$. Then $\bar{z} \in \mathcal{M}_{\gamma_0}$. Since $\delta_j/(\delta_j + g(z_j)) \leq 1$,

$$L_j = \frac{\delta_j \phi_m(z_j)}{\delta_j + g(z_j)} \leq \phi_m(z_j).$$

Continuity of ϕ_m gives

$$\limsup_{j\to\infty} L_j \leq \lim_{j\to\infty} \phi_m(z_j) = \phi_m(\bar{z}) \leq c_m(\varepsilon).$$

In both cases,

$$\limsup_{\ell\to\infty} L_\ell = \lim_{j\to\infty} L_j \leq c_m(\varepsilon).$$

Because the original sequence $(\rho_j)_{j\geq 1}$ approaching $\gamma_0/2$ from below was arbitrary, the sequential characterization of one-sided limits gives

$$\limsup_{\rho\uparrow\gamma_0/2} \frac{\delta_\rho R_{\mathrm{crit}}(m,\rho;\varepsilon)}{\tau} \leq c_m(\varepsilon).$$

Combining this inequality with the previously established lower bound yields

$$\lim_{\rho\uparrow\gamma_0/2} \frac{(\gamma_0 - 2\rho) R_{\mathrm{crit}}(m,\rho;\varepsilon)}{\tau} = c_m(\varepsilon).$$

Finally, since $c_m(\varepsilon) > 0$,

$$\frac{R_{\mathrm{crit}}(m,\rho;\varepsilon)}{\tau c_m(\varepsilon)/(\gamma_0 - 2\rho)} = \frac{(\gamma_0 - 2\rho) R_{\mathrm{crit}}(m,\rho;\varepsilon)}{\tau c_m(\varepsilon)} \longrightarrow 1.$$

Therefore,

$$R_{\mathrm{crit}}(m,\rho;\varepsilon) \sim \frac{\tau c_m(\varepsilon)}{\gamma_0 - 2\rho}$$

as $\rho \uparrow \gamma_0/2$. This proves part *(iii)* and completes the proof. □

H Related Work

Empirical origin and internal mechanisms. Empirical studies have examined when sinks emerge and which model components support them. Gu et al. (2025) found sink behavior across model scales and input types and related its emergence to optimization, data distribution, training objectives, and softmax normalization. Sun et al. (2024) connected attention concentration to massive activations that act as approximately input-independent biases, while Cancedda (2024) associated sinking with low-energy spectral components of intermediate representations. Guo et al. (2024) proposed an active–dormant mechanism linking sink tokens, small value states, and large residual states, together with a theoretical analysis of its training dynamics. At the level of a particular architecture, Ran-Milo et al. (2026) identified a GPT-2 sink circuit involving query bias, positional processing, and the key projection. These results explain how sinks can emerge or be implemented.

Functions and theoretical interpretations. Several works investigate why trained models use sinks. Barbero et al. (2025) argued that attention to an initial token allows a head to limit excessive information mixing when no content token should be selected. Zhang et al. (2026) proposed a catch–tag–release mechanism in which sink-related features organize information for later retrieval. Queipo-de Llano et al. (2026) connected sinks and compression valleys through massive activations and analyzed their role in layerwise information compression. Ruscio et al. (2026) interpreted sinks as reference points in representation space and studied their architecture-dependent topological and spectral structure. Theoretical work has also studied their functional necessity: Ran-Milo (2026) proved that softmax attention must develop a sink to implement a class of trigger-conditional computations, whereas non-normalized attention can solve the same tasks without one. Related analysis by Hong & Lee (2025) connects softmax concentration and entropy collapse to logit variance and training instability. These studies primarily concern sink emergence, function, training, or representation-level organization.

Mitigation, context growth, and quantization. Other work modifies attention to suppress or control sink behavior. Gated attention introduces a mechanism for rejecting uninformative attention outputs and substantially reduces sinks (Qiu et al., 2026), while rectified softmax removes the sum-to-one constraint and eliminates sinks in the studied models (Zuhri et al., 2026). Rather than removing sinks, Son et al. (2024) construct sink-like prefix caches to reduce activation outliers and improve post-training quantization. These methods demonstrate that sink behavior is sensitive to attention normalization and cache design.

KV-cache quantization provides a separate source of perturbation. KIVI uses asymmetric low-bit quantization for cached keys and values (Liu et al., 2024), whereas KVQuant combines per-channel key quantization, non-uniform representations, and explicit outlier handling (Hooper et al., 2024). Their reconstruction errors perturb the key vectors entering attention, but existing accuracy and reconstruction metrics do not yield a head-specific certificate for preserving a sink. Context length creates another difficulty: standard softmax changes as its number of inputs grows, motivating length-dependent normalizations such as scalable softmax (Nakanishi, 2025). Such analyses establish that multiplicity matters, but do not identify its exact interaction with geometric key separation.