Attention as a Laplace-Radon Transform: Tomographic Geometry, Null Spaces, and Kernel Generalizations

[Lawrence P.]

December 18, 2025

Abstract

Scaled dot-product attention is typically presented as a normalized exponential weighting over pairwise query-key dot products. This paper compiles a transform-theoretic representation of a single attention head by showing that the softmax normalizer and value-weighted numerator are exact Laplace transforms of Radon projections of discrete measures on key space. In this representation, each query determines a projection direction (a tomographic "tilt") and a radial Laplace parameter (a "depth" or effective temperature). The factorization yields a geometric diagnostic toolkit for attention heads: (i) per-query tilt and radial statistics; (ii) three exact null spaces acting as information bottlenecks (parameter-level invisibility, dataset limited-angle tomography nulls, and forward-pass routing/value nulls); and (iii) an inverse pipeline for the unnormalized objects via inverse Laplace followed by inverse Radon. Finally, by replacing the Laplace kernel with other positive kernels, attention can be placed in a broader family of generalized Radon integral transforms, enabling principled bias-variance tradeoffs and importing tools from harmonic analysis.

Contents

1 Introduction

Attention mechanisms have become the central primitive in modern sequence models, yet they are often analyzed primarily through optimization dynamics, expressivity results, or empirical interpretability. A complementary viewpoint is to treat attention as an operator acting on distributions of keys and values. This paper compiles a theory in which softmax attention is written (i) as a Laplace transform of a 1D Radon projection (a tomographic "ray transform"), (ii) as a Gibbs measure and an exponential family, and (iii) as a moment-generating function (MGF) of a projected key coordinate.

This transform picture is exact for standard (unmasked) scaled dot-product attention and yields direct geometric characterizations: queries define measurement directions (tilts) on the key sphere; query norms set Laplace radii; the head parameters transport token-space measures into key space; and several linear-algebraic and tomographic null spaces quantify information that the head cannot sense or cannot transmit.

Contributions (compiled from the source theory). The paper organizes the theory into the following contributions.

1. An exact Laplace-Radon factorization of a single attention head.
2. A tomographic interpretation of per-query "tilt" and "radial" parameters, including a head-level bilinear view via $B = W_Q W_K^\top$.
3. Three exact null spaces describing attention bottlenecks: parameter-level nulls ($\ker(W_Q^\top), \ker(W_K^\top), \ker(B)$), dataset limited-angle nulls (tilt-span complements), and forward-pass routing/value nulls ($\ker(P)$).
4. An inverse-transform pipeline for the unnormalized partition and numerator via inverse Laplace and inverse Radon, and a discussion of why normalization blocks a direct inverse from outputs alone.
5. A generalized kernel formulation placing attention in a broader family of generalized Radon integral transforms, including Gibbs/exponential-family identities and the MGF/CGF viewpoint.

2 Background and positioning

The Laplace-Radon factorization makes attention comparable to classical integral transforms used in inverse problems and signal processing. The Radon transform is the foundational operator in computerized tomography and admits well-studied inversion formulas and limited-angle failure

modes. The Laplace transform (and, probabilistically, the moment-generating function) is a classical tool for characterizing measures through their moments and for formulating uniqueness and inversion results.

On the machine learning side, softmax attention originates in Transformers, while fast approximations to exponential dot-product kernels motivate kernel and random-feature interpretations (e.g., Performer-style linear attention). Random Fourier features provide a general approximation framework for shift-invariant kernels and have structured variants with improved variance properties. When attention is restricted to angular structure on the sphere (e.g., dot-product kernels depending only on $u^\top v$), harmonic analysis on $\mathbb{S}^{d-1}$ and classical results on positive definite zonal kernels (Schoenberg) provide an analogue of the Fourier toolkit that benefited convolutional models.

The purpose of this paper is not to propose a new attention mechanism, but to compile an exact and diagnostic transform representation of standard softmax attention and to organize its implications. In particular, the factorization separates:

1. angular structure (which directions are probed),
2. radial structure (how sharply projections are weighted), and
3. the information bottlenecks induced by hard nulls, limited-angle sampling, and routing collapse.

3 Preliminaries and notation

3.1 Scaled dot-product attention

Consider a single attention head with token representations $x_i \in \mathbb{R}^{d_{\text{model}}}$. Let d_k denote the head key/query dimension and write $d := d_k$ for brevity. The standard linear maps are

$$q_j = W_Q^\top x_j, \qquad k_i = W_K^\top x_i, \qquad v_i = W_V^\top x_i. \tag{1}$$

Define scaled logits

$$\ell_{ji} = \frac{q_j^\top k_i}{\sqrt{d}}, \tag{2}$$

and softmax weights and output

$$p_{ji} = \frac{e^{\ell_{ji}}}{\sum_m e^{\ell_{jm}}}, \qquad y_j = \sum_i p_{ji} v_i. \tag{3}$$

This is the standard form of scaled dot-product attention.

3.2 Measures on key space

Fix a sequence (or a forward pass) with keys $\{k_i\}_{i=1}^n \subset \mathbb{R}^{d_k}$ and values $\{v_i\}_{i=1}^n \subset \mathbb{R}^{d_v}$. Define a discrete scalar measure on key space and a vector-valued key-value measure:

$$\mu = \sum_{i=1}^n \delta(\cdot - k_i), \tag{4}$$

$$\nu = \sum_{i=1}^n v_i \, \delta(\cdot - k_i). \tag{5}$$

These encode the empirical key cloud and key-associated value vectors.

Pushforward (measure transport). Given a measurable map $T : X \to Y$ and a measure ρ on X, the pushforward $T_{\#}\rho$ is defined by $(T_{\#}\rho)(A) = \rho(T^{-1}(A))$ for measurable $A \subseteq Y$. In the discrete case $\rho = \sum_i \delta(\cdot - x_i)$, one has $T_{\#}\rho = \sum_i \delta(\cdot - T(x_i))$. This notation makes it explicit that the head parameters $W_K^\top$ transport token-space mass into key space.

3.3 Radon transform of discrete measures

Let $d = d_k$. For $u \in \mathbb{S}^{d-1}$, define the Radon transform of a measure as the pushforward onto the 1D coordinate $s = u^\top k$:

$$(\mathcal{R}\mu)(u, s) = \sum_{i=1}^{n} \delta\big(s - u^\top k_i\big), \tag{6}$$

$$(\mathcal{R}\nu)(u, s) = \sum_{i=1}^{n} v_i \, \delta\big(s - u^\top k_i\big). \tag{7}$$

This aligns with the standard tomographic viewpoint of integrating (or, in the discrete case, accumulating mass) over hyperplanes normal to u.

3.4 Laplace transform and MGF

For a (signed or vector-valued) measure on $\mathbb{R}$ satisfying suitable exponential-moment conditions, the bilateral Laplace transform in variable s evaluated at τ takes the form

$$\mathcal{L}[f](\tau) = \int_{\mathbb{R}} e^{\tau s} f(s) \, \mathrm{d}s, \tag{8}$$

with a corresponding inverse transform under standard hypotheses. When f is a probability measure, this is also the moment-generating function (MGF).

4 Softmax attention as a Laplace-Radon transform

4.1 Query direction and radial parameter

For a query $q_j \in \mathbb{R}^d$, write

$$q_j = \|q_j\| \, u_j, \qquad u_j = \frac{q_j}{\|q_j\|} \in \mathbb{S}^{d-1}, \qquad \tau_j = \frac{\|q_j\|}{\sqrt{d}}. \tag{9}$$

Then the logit becomes $\ell_{ji} = \tau_j \, u_j^\top k_i$.

4.2 Tomographic hyperplanes and the projected coordinate

For a fixed direction $u \in \mathbb{S}^{d-1}$ and coordinate $s \in \mathbb{R}$, define the hyperplane

$$H_{u,s} := \{k \in \mathbb{R}^d : \; u^\top k = s\}. \tag{10}$$

In the continuous setting, the Radon transform can be interpreted as integrating a density over these hyperplanes. In the present discrete-measure setting, $\mathcal{R}\mu$ is a pushforward onto the scalar coordinate $s = u^\top k$ and produces a sum of Dirac deltas at the projected locations. This is the sense in which each query induces a 1D "tomographic view" of the key set.

4.3 Partition function and numerator as Laplace transforms

Define the softmax normalizer (partition function)

$$Z(q_j) = \sum_{i=1}^{n} e^{q_j^\top k_i/\sqrt{d}} = \sum_{i=1}^{n} e^{\tau_j u_j^\top k_i}. \tag{11}$$

By definition of the Radon projection $(\mathcal{R}\mu)(u, s)$, this can be written exactly as

$$Z(q_j) = \int_{\mathbb{R}} e^{\tau_j s} (\mathcal{R}\mu)(u_j, s)\ \mathrm{d}s. \tag{12}$$

Similarly the unnormalized value-weighted sum is

$$N(q_j) = \sum_{i=1}^{n} e^{\tau_j u_j^\top k_i} v_i = \int_{\mathbb{R}} e^{\tau_j s} (\mathcal{R}\nu)(u_j, s)\ \mathrm{d}s. \tag{13}$$

Theorem 1 (Exact Laplace-Radon factorization)**.** *For an unmasked attention head, the output at position j is*

$$\boxed{y_j = \frac{N(q_j)}{Z(q_j)} = \frac{\int_{\mathbb{R}} e^{\tau_j s} (\mathcal{R}\nu)(u_j, s)\ \mathrm{d}s}{\int_{\mathbb{R}} e^{\tau_j s} (\mathcal{R}\mu)(u_j, s)\ \mathrm{d}s}} \tag{14}$$

with $u_j = q_j/\|q_j\|$ and $\tau_j = \|q_j\|/\sqrt{d}$.

Proof. Substitute $q_j = \|q_j\| u_j$ into $q_j^\top k_i/\sqrt{d} = \tau_j u_j^\top k_i$ and rewrite the discrete sums as integrals against the corresponding Radon pushforwards. □

Interpretation. Each query induces a 1D Radon projection of the key(-value) measures along direction u_j, then applies a Laplace transform in the projected coordinate s, and finally normalizes. This makes attention an instance of a normalized integral transform acting on measures.

5 Tomographic geometry: tilt, radius, and head-level structure

5.1 Per-query tilt

In tomography, tilt corresponds to the projection direction. Equation (14) shows that, for query j, the exact tilt is

$$\boxed{u_j = \frac{q_j}{\|q_j\|} \in \mathbb{S}^{d-1}.} \tag{15}$$

All logits depend on each key only through the scalar coordinate $s_{ji} = u_j^\top k_i$ (up to scaling by τ_j).

5.2 Radial (Laplace) parameter

The "radial" parameter in Laplace space is

$$\boxed{\tau_j = \frac{\|q_j\|}{\sqrt{d}}.} \tag{16}$$

Larger τ_j makes the weighting $e^{\tau_j s}$ more extreme along the projected coordinate, whereas small τ_j approaches a flatter weighting.

5.3 Head-level bilinear form and singular geometry

In token space, logits admit an exact bilinear form

$$\ell_{ji} = \frac{x_j^\top B x_i}{\sqrt{d}}, \qquad B := W_Q W_K^\top. \tag{17}$$

A singular value decomposition $B = U\Sigma V^\top$ provides a diagnostic description: right singular vectors V identify token-feature directions that affect keys, left singular vectors U identify token-feature directions that affect queries, and singular values encode gains. Rapid spectral decay corresponds to a head that effectively probes a limited set of paired directions, analogous to limited-angle scanning.

5.4 Tilt structure induced by parameters

The decomposition $B = U\Sigma V^\top$ makes "global" interaction directions explicit. One convenient summary is the leading singular vectors u_1 and v_1, which identify dominant query-side and key-side directions. When singular values decay quickly, the head effectively concentrates its measurements in a low-dimensional interaction subspace, paralleling "few-view" behavior in tomography.

6 Null spaces as information bottlenecks

The transform viewpoint makes several "nulls" precise. They are exact linear-algebraic statements, but the tomographic picture clarifies their geometric meaning.

6.1 Parameter-level nulls (hard invisibility)

These null spaces are independent of data. If $z \in \ker(W_K^\top)$, then $W_K^\top(x_i + \alpha z) = W_K^\top x_i$ for all α, so the head cannot sense that feature direction through keys. Similarly $\ker(W_Q^\top)$ consists of query-side invisible directions. In the bilinear form representation,

$$\ker(B) = \{z : Bz = 0\}, \qquad \ker(B^\top) = \{z : B^\top z = 0\} \tag{18}$$

give key-side and query-side logit nulls, respectively.

Exact operator statement. The hard key-side invisibility subspace and query-side invisibility subspace are

$$\ker(W_K^\top) = \{z \in \mathbb{R}^{d_{\text{model}}} : W_K^\top z = 0\}, \qquad \ker(W_Q^\top) = \{z \in \mathbb{R}^{d_{\text{model}}} : W_Q^\top z = 0\}. \tag{19}$$

These directions are destroyed before any tomographic projection occurs. The bilinear nulls $\ker(B)$ and $\ker(B^\top)$ are coarser invariances at the logit level: if $z \in \ker(B)$ then $x_j^\top B(x_i + \alpha z) = x_j^\top B x_i$ for all x_j and α, so logits are unchanged for every query.

6.2 Dataset limited-angle nulls (tomographic missing views)

Even if the parameter maps are full rank, the head may only realize a limited set of tilts on a dataset. Let

$$U := \{u_j\}_{j,\text{examples}} \subset \mathbb{S}^{d-1} \tag{20}$$

be the set of tilts observed across positions/examples. Then any key-space direction n orthogonal to $\mathrm{span}(U)$ is never probed:

$$\boxed{\mathcal{N}_{\text{tilt,key}} = (\mathrm{span}(U))^{\perp}.} \tag{21}$$

This is the exact analogue of limited-angle tomography ("missing wedge").

Angular coverage as a diagnostic. The quantity $\dim \mathrm{span}(U)$ (or, more robustly, the spectrum of the empirical covariance $\mathbb{E}[uu^{\top}]$) measures how many independent angular directions a head actually scans on data. Limited-angle effects arise when $\mathrm{span}(U)$ is low-dimensional relative to $\mathbb{R}^d$, leaving a large orthogonal complement that is never probed.

The corresponding token-space measurement normals are

$$N := \{n_j\}_{j,\text{examples}}, \qquad n_j := W_K u_j. \tag{22}$$

Thus the token-space limited-angle null is

$$\boxed{\mathcal{N}_{\text{tilt,token}} = (\mathrm{span}(N))^{\perp}.} \tag{23}$$

6.3 Forward-pass routing/value nulls

For a fixed forward pass, stack the attention weights into $P \in \mathbb{R}^{n_q \times n_k}$ with entries p_{ji}. Then for stacked values $V \in \mathbb{R}^{n_k \times d_v}$ the head outputs satisfy

$$Y = PV. \tag{24}$$

The exact null space in value space is

$$\boxed{\mathcal{N}_V = \{\Delta V \in \mathbb{R}^{n_k \times d_v} : P\Delta V = 0\}.} \tag{25}$$

Equivalently, per value dimension, this is $\ker(P)$. An SVD $P = U\Sigma R^{\top}$ yields $\ker(P)$ as the span of right singular vectors corresponding to zero singular values. Geometrically, even if a direction is representable by W_V, the current routing P may project it out.

Dimension of the routing null. Since $\dim \ker(P) = n_k - \mathrm{rank}(P)$, row-wise saturation of P (e.g., nearly one-hot attention) can reduce $\mathrm{rank}(P)$ and enlarge the routing null space. This links "sharp" attention regimes to value-space information loss.

6.4 Summary: exact geometric objects per head

For reference, the key objects and null spaces compiled in this paper admit the following equation-first summary:

$$\boxed{y_j = \frac{\int_{\mathbb{R}} e^{\tau_j s}(\mathcal{R}\nu)(u_j, s)\ \mathrm{d}s}{\int_{\mathbb{R}} e^{\tau_j s}(\mathcal{R}\mu)(u_j, s)\ \mathrm{d}s}}, \qquad \boxed{u_j = \frac{q_j}{\|q_j\|}}, \qquad \boxed{\tau_j = \frac{\|q_j\|}{\sqrt{d}}}. \tag{26}$$

$$\boxed{\mathcal{N}_{\text{tilt,key}} = (\mathrm{span}\{u_j\})^{\perp}}, \qquad \boxed{\mathcal{N}_V = \ker(P)}, \qquad \boxed{\ker(W_K^{\top}),\ \ker(W_Q^{\top}),\ \ker(B)}\ \text{(hard parameter nulls)}. \tag{27}$$

7 Radial extension, conditioning, and Monte Carlo approximation

The Laplace-Radon form separates two kinds of coverage: (i) angular coverage of tilts u, and (ii) radial coverage of Laplace parameters τ. A head has less tomographic null space when it increases the effective dimension of $\mathrm{span}(U)$ and broadens the distribution of τ.

7.1 Angular extension (more tilts)

Angular extension corresponds to a larger $\dim \mathrm{span}(U)$ on data. At the parameter level, this is promoted when the head bilinear form $B = W_Q W_K^\top$ has higher effective rank and data excites multiple singular directions.

7.2 Radial extension (more τ support)

Radial extension corresponds to a broader distribution of

$$\tau_j = \frac{\|q_j\|}{\sqrt{d}} = \frac{\|W_Q^\top x_j\|}{\sqrt{d}} \tag{28}$$

and a broader spread of projected coordinates $s = u^\top k$. Multiple τ values probe different moments of the Radon projection. If τ is essentially fixed and small, the transform becomes low-contrast along s and different projected distributions can become difficult to distinguish.

7.3 Null-space interference in Monte Carlo bounds

Monte Carlo (MC) approximations of attention often target

$$Z(q) = \sum_i e^{q^\top k_i/\sqrt{d}}, \qquad N(q) = \sum_i e^{q^\top k_i/\sqrt{d}} v_i, \tag{29}$$

or kernelized equivalents, by sampling or random features (e.g., Performer-style approximations). In this setting, "null-space interference" occurs when the estimator fails to sense directions/components that materially affect Z and N, increasing variance or bias.

The theory compiles several ways to extend radial/angle coverage and reduce such interference.

1. *Head-wise temperature / logit scaling.* A learnable scalar α_h rescales logits

$$\ell_{ji} = \alpha_h \frac{q_j^\top k_i}{\sqrt{d}}, \tag{30}$$

   which directly maps $\tau_j \mapsto \alpha_h \tau_j$. Excessively large scaling can saturate P toward one-hot rows, reducing $\mathrm{rank}(P)$ and enlarging $\ker(P)$, so the goal is broader τ support without rank collapse.

2. *Multi-head angular diversity.* Multiple heads provide multiple tilt families. Encouraging $B_h = W_{Q,h} W_{K,h}^\top$ to span distinct singular directions across heads increases angular coverage.

3. *MC feature/ray coverage.* In random-feature approximations of exponential kernels, finite feature dimension induces an approximation null. Increasing the number of features and using structured or orthogonal features improves directional coverage and reduces variance at fixed dimension.

4. *Importance sampling aligned to $e^{\tau s}$.* Since the integrand scales as $e^{\tau s}$, naive uniform sampling over keys can have high variance when τ is large. Proposals aligned to the distribution of $s = u^\top k$ reduce variance.

8 Parameters as geometric transforms

Beyond describing outputs, the transform formulation locates the parameters as the learned geometry.

8.1 Parameters as pushforwards of measures

Define a discrete token measure $\rho = \sum_i \delta(\cdot - x_i)$. Then $W_K^\top$ transports token mass into key space: $\mu = (W_K^\top)_\# \rho$. Thus the "object" on which tomography is performed is itself produced by a learned linear change of variables.

8.2 Tilt pullback into token space

For any $u \in \mathbb{S}^{d-1}$,

$$u^\top k_i = u^\top W_K^\top x_i = (W_K u)^\top x_i. \tag{31}$$

Therefore Radon projection in key space along u corresponds to projection in token space along the normal $W_K u$. For query j, with $u_j = q_j/\|q_j\|$, the induced token-space normal is

$$\boxed{n_j := W_K u_j = W_K \frac{W_Q^\top x_j}{\|W_Q^\top x_j\|}.} \tag{32}$$

Logits can be written $\ell_{ji} = \tau_j n_j^\top x_i$ with $\tau_j = \|W_Q^\top x_j\|/\sqrt{d}$. This provides an exact data-dependent "scan direction" in token space.

8.3 An L2 change-of-variables interpretation

The bilinear form $B = W_Q W_K^\top$ is a general pairing. If one restricts to a symmetric positive semidefinite pairing $G \succeq 0$ (or analyzes a PSD surrogate), then $G = A^\top A$ yields

$$x_j^\top G x_i = (Ax_j)^\top (Ax_i), \tag{33}$$

an explicit Euclidean (L2) change of coordinates prior to a dot product.

9 Inverse problems and identifiability

9.1 Inverse transform for unnormalized objects

The Laplace-Radon representation provides a literal inverse pipeline for the unnormalized quantities. For each direction u, define

$$Z(u,\tau) = \int_{\mathbb{R}} e^{\tau s} (\mathcal{R}\mu)(u,s)\, \mathrm{d}s, \qquad N(u,\tau) = \int_{\mathbb{R}} e^{\tau s} (\mathcal{R}\nu)(u,s)\, \mathrm{d}s. \tag{34}$$

For fixed u, $Z(u,\cdot)$ is a Laplace transform of $(\mathcal{R}\mu)(u,\cdot)$. Formally, the inverse Laplace transform (Bromwich inversion) recovers the projected measure:

$$(\mathcal{R}\mu)(u,s) = \mathcal{L}^{-1}_{\tau\to s}[Z(u,\tau)] = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{-\tau s} Z(u,\tau)\, \mathrm{d}\tau. \tag{35}$$

An analogous inversion holds componentwise for $(\mathcal{R}\nu)(u,s)$ from $N(u,\tau)$. Once Radon data is available for all u and s, classical Radon inversion (e.g., filtered backprojection written in operator form) recovers μ and ν in an appropriate distributional sense.

9.2 Function classes and injectivity of the forward transform

There are two complementary "faithfulness" questions. The first asks when the transform $\mu \mapsto Z$ is injective (i.e., whether the underlying key measure can be recovered from unnormalized observations). The second asks what can be recovered from the normalized output map $(Q, K, V) \mapsto Y$.

Transform level ($\mu \mapsto Z$). For each fixed direction u, the quantity $Z(u, \tau)$ is the Laplace transform of the projected measure $(\mathcal{R}\mu)(u, \cdot)$. Under standard exponential-moment conditions ensuring existence of the Laplace transform on an interval or strip in τ, uniqueness theorems imply injectivity: $Z(u, \tau)$ determines $(\mathcal{R}\mu)(u, s)$ as a distribution. If $(\mathcal{R}\mu)(u, s)$ is available for all directions $u \in \mathbb{S}^{d-1}$, Radon inversion is injective (in an appropriate distributional sense), yielding recovery of μ.

Discrete case (exponential sums). When keys are discrete, for fixed u one has

$$Z(u, \tau) = \sum_{i=1}^{n} e^{\tau s_i}, \qquad s_i = u^\top k_i, \tag{36}$$

an exponential sum ("exponential polynomial"). Knowing this function on a τ-interval probes the moment sequence of the projection and, under mild nondegeneracy, identifies the multiset $\{s_i\}$. Thus, radial coverage in τ has a concrete interpretation as access to higher-order projected moments.

Network level ($(Q, K, V) \mapsto Y$). The normalized output $y = N/Z$ generally does not uniquely identify Z and N. Inversion is therefore limited unless additional observables are available (e.g., logits, attention weights, or the ability to probe the same (Q, K) with multiple value assignments V to recover the routing matrix P).

9.3 Why softmax normalization blocks a direct inverse

The head output is the ratio

$$y(u, \tau) = \frac{N(u, \tau)}{Z(u, \tau)}. \tag{37}$$

In general, this ratio does not determine Z and N uniquely: multiplying both by the same positive scalar function of (u, τ) leaves y unchanged. More broadly, many different base measures can induce the same conditional expectation. Thus inversion is possible only if one can access sufficient unnormalized information (e.g., logits or attention weights plus a scale anchor).

9.4 Identifiability of parameters and gauge symmetry

Given enough constraints of the form

$$\ell_{ji} = \frac{x_j^\top B x_i}{\sqrt{d}}, \qquad B = W_Q W_K^\top, \tag{38}$$

the bilinear form B is the identifiable geometric object. However, factoring B into W_Q and W_K is not unique: for any invertible $A \in \mathbb{R}^{d \times d}$,

$$W_Q \to W_Q A, \qquad W_K \to W_K A^{-\top} \tag{39}$$

leaves B unchanged. This is a parameter "gauge symmetry" that implies factor-level inversion requires additional conventions or constraints. Under a PSD metric restriction $G = A^\top A$, the factorization is canonical up to orthogonal rotation.

10 Generalized kernels and generalized Radon integral transforms

The Laplace-Radon viewpoint generalizes beyond softmax. Let $\kappa(q,k) \geq 0$ be any positive kernel and define

$$Z_\kappa(q) = \int \kappa(q,k)\ \mathrm{d}\mu(k) = \sum_i \kappa(q,k_i), \tag{40}$$

$$N_\kappa(q) = \int \kappa(q,k)\ \mathrm{d}\nu(k) = \sum_i \kappa(q,k_i) v_i, \tag{41}$$

$$y(q) = \frac{N_\kappa(q)}{Z_\kappa(q)}. \tag{42}$$

10.1 Ridge form and generalized 1D transform

Assume the kernel depends on keys through a 1D projection:

$$q = \|q\| u, \quad u \in \mathbb{S}^{d-1}, \qquad \kappa(q,k) = \psi(\tau, u^\top k), \tag{43}$$

with some radial parameter $\tau = \tau(q)$. Then define the generalized attention transform

$$\boxed{(\mathcal{T}_\psi \mu)(u,\tau) := \int_{\mathbb{R}} \psi(\tau,s)(\mathcal{R}\mu)(u,s)\ \mathrm{d}s.} \tag{44}$$

Softmax corresponds to $\psi(\tau,s) = e^{\tau s}$.

10.2 Gibbs measure identity

For any strictly positive kernel, define an energy $E(q,k) := -\log \kappa(q,k)$. Then the attention weights form a Gibbs/Boltzmann distribution over keys:

$$p(k_i \mid q) = \frac{e^{-E(q,k_i)}}{\sum_m e^{-E(q,k_m)}}. \tag{45}$$

The output is a conditional expectation $y(q) = \mathbb{E}_{p(\cdot|q)}[v]$. This identity clarifies that energies are identifiable only up to adding a q-dependent constant.

10.3 Softmax as an exponential family

Softmax attention uses

$$\kappa(q,k) = \exp\Big(\tfrac{1}{\sqrt{d}} q^\top k\Big). \tag{46}$$

This is an exponential family with sufficient statistic $T(k) = k$, natural parameter $\eta(q) = q/\sqrt{d}$, and log-partition function

$$A(\eta) = \log \sum_m e^{\eta^\top k_m}. \tag{47}$$

Standard exponential-family identities yield

$$\nabla_\eta A(\eta) = \mathbb{E}[k], \qquad \nabla_\eta^2 A(\eta) = \mathrm{Cov}(k), \tag{48}$$

and equivalently in q,

$$\nabla_q \log Z(q) = \frac{1}{\sqrt{d}} \mathbb{E}[k]. \tag{49}$$

10.4 MGF/CGF view of projected coordinates

Fix u and let $S = u^\top K$ where $K \sim \mu$. Let μ_u denote the 1D projected measure. For softmax,

$$Z(u, \tau) = \int e^{\tau s} \, \mu_u(\mathrm{d}s) \tag{50}$$

is exactly the MGF $M_S(\tau) = \mathbb{E}[e^{\tau S}]$. Then $\log Z$ is the cumulant-generating function (CGF), and derivatives yield cumulants, for example

$$\partial_\tau \log Z(u, \tau) = \mathbb{E}[S], \qquad \partial_\tau^2 \log Z(u, \tau) = \mathrm{Var}(S), \tag{51}$$

where the expectation is under the Gibbs-reweighted distribution. Expanding the MGF makes the moment channel explicit:

$$Z(u, \tau) = \sum_{m=0}^{\infty} \frac{\tau^m}{m!} \int s^m \, \mu_u(\mathrm{d}s). \tag{52}$$

10.5 Swapping kernels and bias-variance tradeoffs

Once attention is written as $\mathcal{T}_\psi$ applied to Radon projections, replacing ψ swaps the 1D transform applied to each projection. This allows importing known approximation and inversion behavior. Examples compiled in the source theory include:

1. Exponential/Laplace kernel $\psi(\tau, s) = e^{\tau s}$ (softmax): high dynamic range; invertible in principle via inverse Laplace, but inversion is ill-conditioned under noise.
2. Gaussian-on-projection smoothing kernels: lower-variance estimates but stronger smoothing and a more ill-posed deconvolution-style inverse.
3. Cauchy/Lorentzian smoothing: heavier tails and slower spectral decay than Gaussian, often corresponding to less aggressive smoothing.
4. α-stable families: fractional smoothing tunable by $\alpha \in (0, 2]$, interpolating between Gaussian-like and heavier-tailed behavior.

For shift-invariant kernels on the projection coordinate, random Fourier features provide standard fast estimators, and orthogonal/structured features can reduce variance.

10.6 Worked kernel families on the projection coordinate

The kernel-swap idea can be made explicit by choosing $\psi(\tau, s)$ families in the 1D projection coordinate. The expressions below are organized to emphasize conditioning and estimator design.

Exponential/Laplace kernel (softmax). Softmax corresponds to $\psi(\tau, s) = e^{\tau s}$, so

$$(\mathcal{T}_\psi \mu)(u, \tau) = \int_{\mathbb{R}} e^{\tau s} (\mathcal{R}\mu)(u, s) \, \mathrm{d}s. \tag{53}$$

Inversion proceeds formally by inverse Laplace followed by inverse Radon; Laplace inversion is ill-conditioned under noise.

Gaussian-on-projection (Weierstrass/heat smoothing). A shift-invariant kernel on the projection coordinate is

$$\psi_\sigma(\tau, s) := \exp\Big(-\frac{(\tau - s)^2}{2\sigma^2}\Big). \tag{54}$$

Then $\mathcal{T}_{\psi_\sigma}$ is convolution in the s-coordinate:

$$(\mathcal{T}_{\psi_\sigma}\mu)(u, \tau) = (g_\sigma * (\mathcal{R}\mu)(u, \cdot))(\tau), \tag{55}$$

where $g_\sigma(t) = \exp(-t^2/(2\sigma^2))$. In Fourier over τ, this multiplies by $e^{-\sigma^2\omega^2/2}$; any inverse becomes a deconvolution with multiplier $e^{+\sigma^2\omega^2/2}$, amplifying high-frequency noise.

Cauchy/Lorentzian (heavier tails). A heavier-tailed alternative is

$$\psi_\gamma(\tau, s) := \frac{1}{1 + \big((\tau - s)/\gamma\big)^2}. \tag{56}$$

This is also shift-invariant; its spectral decay is slower than Gaussian, retaining more high-frequency content at a comparable width.

α-stable (fractional diffusion families). A common parameterization is by a Fourier multiplier

$$\widehat{\psi_\alpha}(\omega) = e^{-c|\omega|^\alpha}, \qquad 0 < \alpha \le 2. \tag{57}$$

Here $\alpha = 2$ recovers Gaussian-type smoothing, while smaller α yields heavier tails.

Constraint: probability weights typically require nonnegative kernels. If attention weights are required to be probabilities, one typically enforces $\kappa(q, k) \ge 0$ so that $p_i \propto \kappa(q, k_i)$ defines a valid distribution. This rules out many oscillatory kernels (e.g., pure Fourier characters) unless one moves to signed/complex weights or represents such kernels indirectly.

Estimator notes. When $\psi(\tau, s)$ is shift-invariant (depends only on $\tau - s$), random Fourier feature constructions apply directly on the projection coordinate. For softmax in the original dot-product form, positive random feature schemes such as FAVOR+ yield linear-time estimators.

11 Head dimension, harmonic analysis, and the CNN/Fourier analogy

The transform viewpoint suggests a harmonic analysis route analogous to Fourier analysis for CNNs. Tilts live on $\mathbb{S}^{d-1}$. Dot-product kernels restricted to the sphere define zonal kernels, which diagonalize in spherical harmonics. Positive definite zonal kernels on spheres admit classical characterizations (Schoenberg) that connect kernel choice to spectral coefficient decay and effective bandwidth.

The head dimension $d = d_k$ has three precise roles.

1. It sets the ambient dimension of the angular manifold $\mathbb{S}^{d-1}$: as d increases, full angular coverage requires more diverse tilts.

2. It bounds the interaction rank: $\mathrm{rank}(W_Q W_K^\top) \le d$.

3. Through the $1/\sqrt{d}$ scaling, it affects the effective temperature $\tau = \|q\|/\sqrt{d}$ and thus the sharpness of the exponential kernel.

In this picture, "angular resolution" is controlled by realized tilt diversity (data and parameters), while "bandwidth" is controlled by kernel choice and temperature statistics.

12 Implications and diagnostic predictions

The compiled theory is diagnostic rather than merely metaphorical: the Laplace-Radon factorization is exact and exposes measurable geometric quantities. Implications include:

1. *Tilt diagnostics.* For a dataset, estimate the distribution of tilts u_j per head and the dimension/spectrum of $\mathrm{span}(U)$ as a measure of limited-angle behavior.

2. *Radial diagnostics.* Track the distribution of τ_j and the resulting sharpness/entropy of attention rows; identify regimes where increased τ causes routing rank collapse.

3. *Null-space diagnostics.* Compute $\ker(W_Q^\top), \ker(W_K^\top)$ (parameter-level) and $\ker(P)$ per forward pass (routing-level) to detect information bottlenecks.

4. *MC approximation design.* Use the transform picture to align sampling proposals and random-feature constructions with the dominant $s = u^\top k$ coordinates and the effective τ regime.

5. *Kernel exploration.* Replace the Laplace kernel by other positive kernels to trade inversion conditioning against smoothing, leveraging existing approximation theory.

12.1 Concrete diagnostic quantities

The transform picture suggests a small set of summaries that can be computed directly from a model and a dataset. For a head with per-token tilts $u_j = q_j/\|q_j\|$ and radii $\tau_j = \|q_j\|/\sqrt{d}$, define the empirical tilt covariance

$$C_u := \mathbb{E}[uu^\top]. \tag{58}$$

Its eigenvalue spectrum indicates the effective angular dimensionality of scanning directions on the dataset. Likewise, the empirical distribution of τ quantifies the "radial" regime (flat vs sharp weighting). For a given forward pass with attention matrix P, the singular spectrum of P and the dimension $\dim\ker(P) = n_k - \mathrm{rank}(P)$ quantify routing collapse. These diagnostics jointly distinguish: (i) hard parameter nulls, (ii) limited-angle (data-dependent) nulls, and (iii) routing (forward-pass) nulls.

13 Empirical validation: synthetic attention tomography experiments

This section reports empirical checks of the identities and diagnostics in the transform picture. All plots and numbers come from the synthetic run stored in `runs/run_20251217_190709`. The experiments are implemented by `experiment_attention_tomography.py` (data generation, identity/gauge checks, and diagnostic plots) and `compare_theory_vs_empirical.py` (distributional overlays and mean-field baselines).

Quantity	Value (run_20251217_190709)
d_{model} / d_k / d_v	96 / 32 / 48
Tokens per example n	64
Examples	256
Kernel	softmax
$\text{Var}(q)$ per dim / $\text{Var}(k)$ per dim	1.0706 / 0.9665
Predicted logit std $\sqrt{\text{Var}(q)\text{Var}(k)}$	1.0172
Empirical logit std	1.0125
τ mean (empirical / chi prediction)	1.0242 / 1.0267
Laplace-Radon rel. error $\|Y - Y_{\text{hat}}\|/\|Y\|$	5.94×10^{-16}
Projection logit rel. error	2.89×10^{-16}
Gauge rel. error (B / logits / Y)	2.31×10^{-15} / 3.11×10^{-15} / 3.03×10^{-15}
$\dim_{\text{eff}} \text{span}(U)$ / $\dim_{\text{eff}} \text{span}(N)$	24.53 / 19.93
Effective rank(P) at $\alpha = 1$	37.14

Table 1: Summary of the synthetic experiment run used for empirical validation.

13.1 Experimental setup

We isolate a single attention head on synthetic token batches. Tokens $x_i \in \mathbb{R}^{d_{\text{model}}}$ are sampled i.i.d. from a standard Gaussian. A single head is instantiated by random Gaussian matrices W_Q, W_K, W_V (scaled by $1/\sqrt{d_{\text{model}}}$), producing

$$q_j = W_Q^\top x_j, \qquad k_i = W_K^\top x_i, \qquad v_i = W_V^\top x_i. \tag{59}$$

With $d_k = 32$, $d_{\text{model}} = 96$, $d_v = 48$, and $n = 64$ tokens per example, the experiment computes:

1. the exact Laplace-Radon reconstruction Y_{hat} from (u, τ, K, V),
2. projection-only dependence of logits ($\ell_{ji} = \tau_j u_j^\top k_i$),
3. gauge invariance under $W_Q \to W_Q A$, $W_K \to W_K A^{-\top}$,
4. limited-angle diagnostics via spectra of $\text{Cov}(u)$ and $\text{Cov}(n)$ with $n = W_K u$,
5. routing spectra of P and a temperature sweep $\ell \mapsto \alpha\ell$.

13.2 Exact Laplace-Radon identity and 1D dependence

For the softmax kernel, the transform theory predicts the exact identity

$$y_j = \frac{\sum_i e^{\tau_j u_j^\top k_i} v_i}{\sum_i e^{\tau_j u_j^\top k_i}}, \qquad u_j = \frac{q_j}{\|q_j\|}, \quad \tau_j = \frac{\|q_j\|}{\sqrt{d_k}}. \tag{60}$$

The experiment reconstructs Y_{hat} from (u, τ, K, V) and verifies $Y \approx Y_{\text{hat}}$ to machine precision (Table 1). It also verifies the projection-only dependence of logits,

$$\ell_{ji} = \frac{q_j^\top k_i}{\sqrt{d_k}} = \tau_j u_j^\top k_i. \tag{61}$$

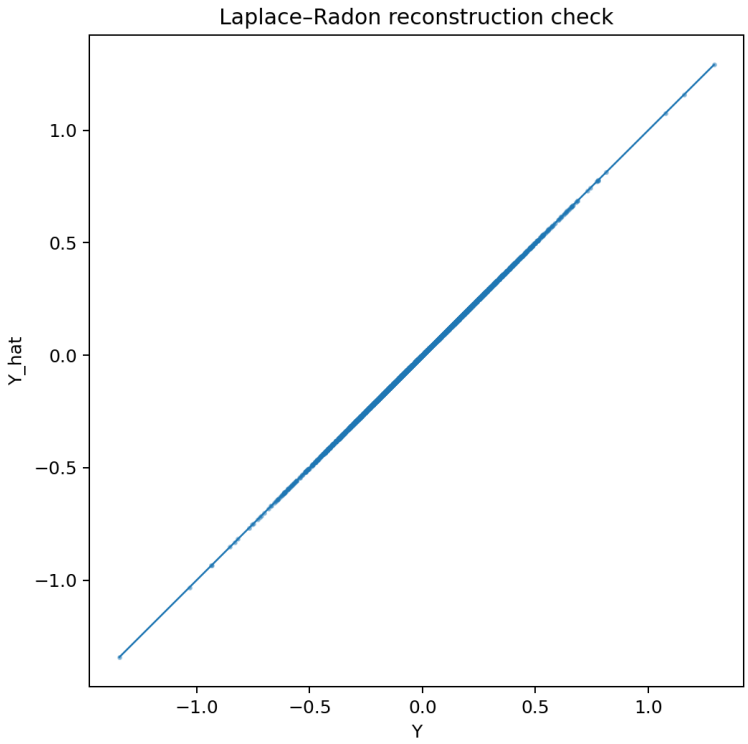

Figure 1: Laplace-Radon reconstruction check: scatter of Y (direct softmax attention) vs Y_{hat} computed from the Laplace-Radon form. The near-perfect diagonal alignment corresponds to a relative error $\|Y - Y_{\text{hat}}\|/\|Y\| \approx 5.94 \times 10^{-16}$.

Implication. The Laplace-Radon factorization is not merely qualitative: for softmax it is an exact algebraic reparameterization. This supports using (u, τ) as literal per-query scan parameters and motivates diagnostics computed directly from these quantities.

13.3 Gauge symmetry check

The factorization $B = W_Q W_K^\top$ is invariant to the gauge transform $W_Q \to W_Q A$, $W_K \to W_K A^{-\top}$. The experiment applies a random invertible A and confirms that B, logits, and outputs are unchanged to numerical precision (Table 1).

Implication. This empirically illustrates that B is the identifiable geometric object, while (W_Q, W_K) are only defined up to gauge.

13.4 Radial parameter distribution and chi prediction

Under the synthetic data model, query vectors are approximately Gaussian with per-dimension variance estimated from the run. The theory therefore predicts that

$$\tau = \frac{\|q\|}{\sqrt{d_k}} \tag{62}$$

follows a scaled chi distribution. Figure 2 shows that the empirical density of τ closely tracks this prediction.

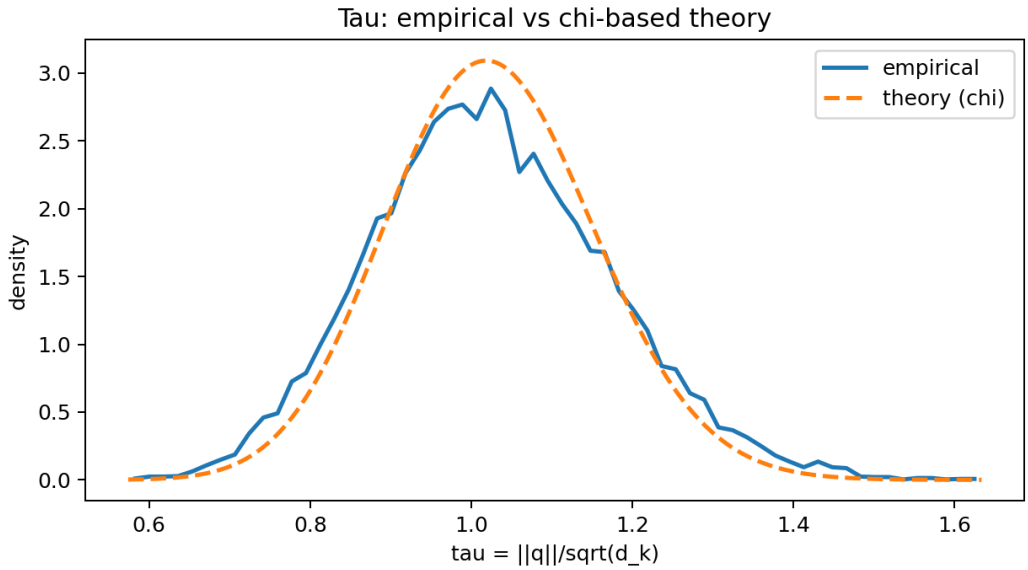

Figure 2: Empirical τ density (blue) vs chi-based prediction (dashed) from the Gaussian mean-field model.

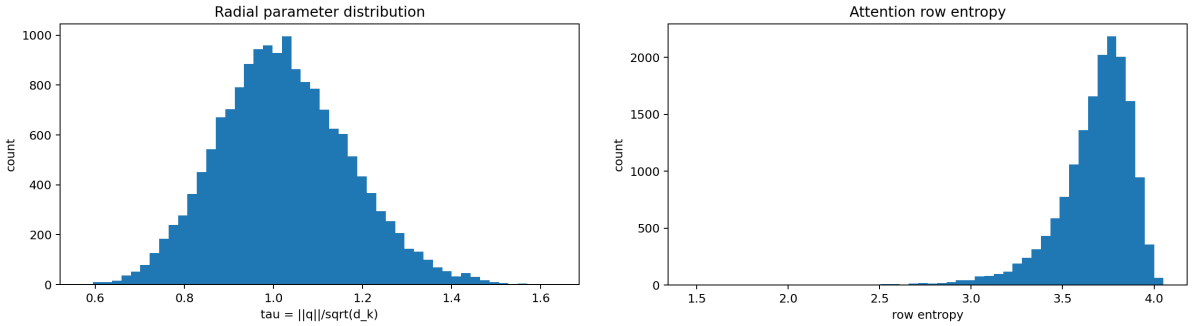

Figure 3: Raw distributions at $\alpha = 1$. Left: histogram of radial parameters $\tau = \|q\|/\sqrt{d_k}$. Right: histogram of attention-row entropies $-\sum_i p_{ji} \log p_{ji}$.

Implication. The radial parameter τ can be modeled and tracked statistically (here, via a chi law), enabling head-wise "temperature" regime classification.

13.5 Logit distribution and Gaussian mean-field prediction

If q and k are treated as independent Gaussians with per-dimension variances $\mathrm{Var}(q)$ and $\mathrm{Var}(k)$, then

$$\ell = \frac{q^\top k}{\sqrt{d_k}} \approx \mathcal{N}\big(0, \mathrm{Var}(q)\mathrm{Var}(k)\big). \tag{63}$$

Figure 4 overlays the empirical logit distribution with this Gaussian prediction.

Implication. Distributional approximations at the logit level can be accurate enough to predict downstream statistics of attention (entropy, max probability, and routing rank) under temperature

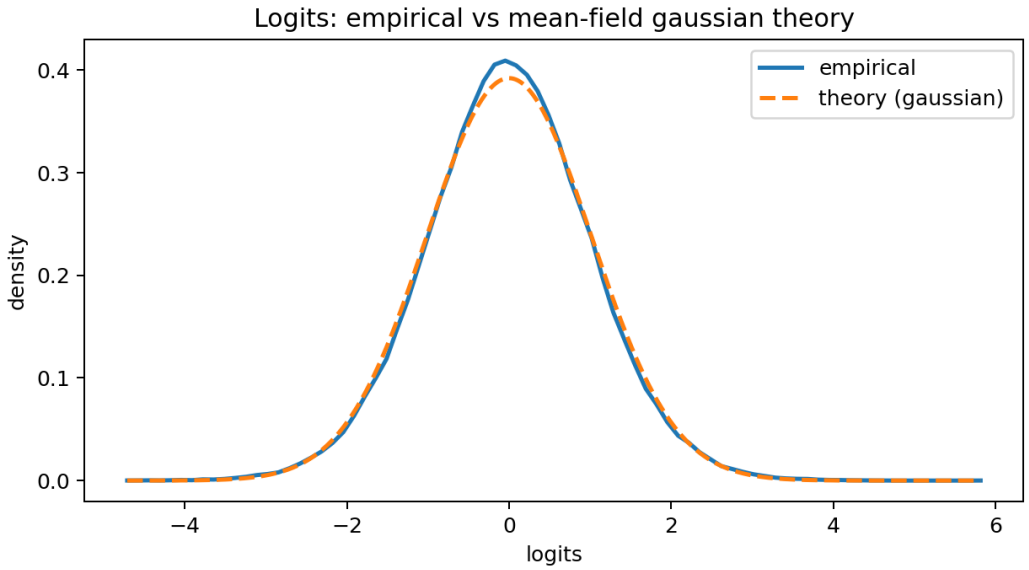

Figure 4: Empirical logits (blue) vs Gaussian mean-field prediction (dashed) using the observed per-dimension variances.

scaling.

13.6 Limited-angle diagnostics: tilt coverage in key and token space

The limited-angle null is controlled by the span of observed tilts u_j. Figure 5 shows the eigenvalue spectrum of the empirical tilt covariance $\mathrm{Cov}(u)$ and reports an effective angular dimension $\dim_{\mathrm{eff}} \mathrm{span}(U) \approx 24.53$. Pushing tilts back into token space via $n = W_K u$ yields the token-space measurement normals; Figure 6 shows the corresponding spectrum and effective dimension $\dim_{\mathrm{eff}} \mathrm{span}(N) \approx 19.93$.

Implication. Even with full-rank parameters (here, $\ker(W_Q^\top) = \ker(W_K^\top) = \{0\}$ in this random initialization), realized tilt diversity can be substantially lower than d_k, and token-space scan directions can be more constrained than key-space tilts.

13.7 Routing spectra and temperature sweep

The routing matrix P acts as a linear map on values; its singular spectrum and effective rank quantify routing collapse. Figure 7 shows the singular values of a representative P at $\alpha = 1$ with effective rank ≈ 37.14.

The theory predicts that scaling logits by α (equivalently, scaling $\tau \mapsto \alpha\tau$) induces a systematic tradeoff: row entropy decreases, max probability increases, and routing rank can peak and then decline as rows saturate. In this run, entropy drops from 4.15 at $\alpha = 0.1$ to 0.48 at $\alpha = 10$, while the mean max row probability rises from 0.0197 to 0.836. The effective rank peaks at $\alpha \approx 2.0$ with $\mathrm{rank}_{\mathrm{eff}}(P) \approx 40.49$. Figures 8-10 show empirical sweeps over $\alpha \in [0.1, 10]$ together with two Monte Carlo theory baselines: (i) i.i.d. Gaussian logits and (ii) structured bilinear logits $L = QK^\top/\sqrt{d_k}$ with Gaussian Q, K.

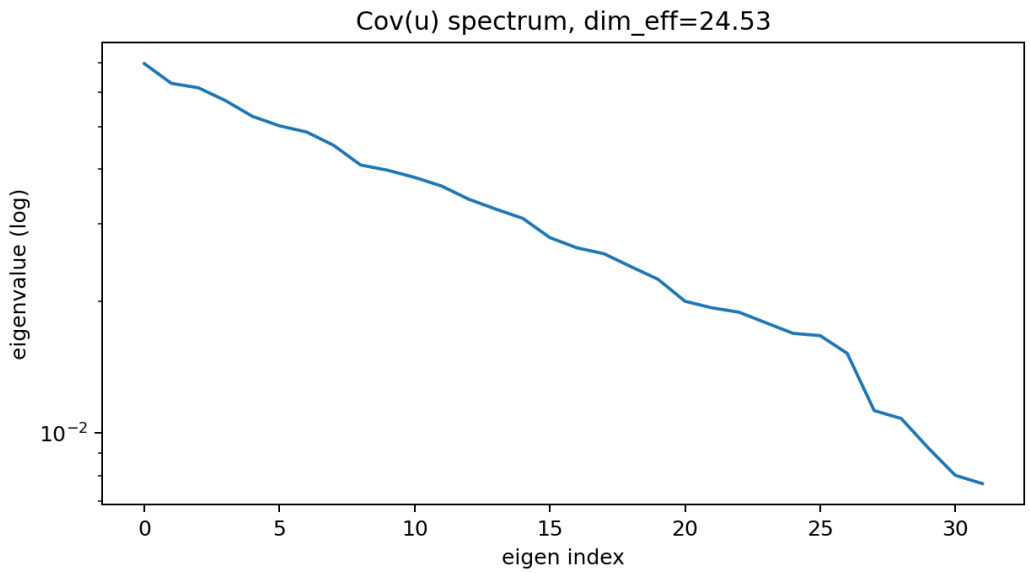

Figure 5: Spectrum of $\mathrm{Cov}(u)$ for per-query tilts $u = q/\|q\|$. Slow decay indicates broad angular coverage; the reported effective dimension is $\dim_{\mathrm{eff}} \mathrm{span}(U) \approx 24.53$.

Implication. The sweep provides an empirical link between the radial parameter distribution (effective temperature) and routing collapse: increasing α pushes the head toward low-entropy, near-deterministic routing, which can enlarge value-space nulls by reducing the effective rank of P.

14 Limitations and open problems

Several limitations are intrinsic. First, the Radon and Laplace inverses are distributional and can be ill-conditioned with noise. Second, the attention output is normalized, and the ratio N/Z generally does not identify the underlying measures without additional observables. Third, masking, positional structure, and full transformer composition (layer norms, MLP nonlinearities, residual connections) introduce additional nonlinearities beyond a single-head transform. Finally, parameter identification from $B = W_Q W_K^\top$ has gauge ambiguity without extra constraints.

Open problems include quantifying the stability of approximate inverses under realistic noise, defining practical "tilt coverage" metrics tied to downstream performance, and developing kernel-swapped attention mechanisms with controlled bias-variance behavior.

15 Conclusion

This paper compiles a transform-theoretic view of attention in which softmax attention is exactly a Laplace transform of Radon projections of discrete key(-value) measures. The representation yields precise geometric diagnostics (tilt, radius) and identifies three exact null spaces that act as information bottlenecks. It also provides an inverse-transform pipeline for unnormalized quantities and situates attention in a broader family of generalized Radon integral transforms, enabling principled kernel design and connections to harmonic analysis.

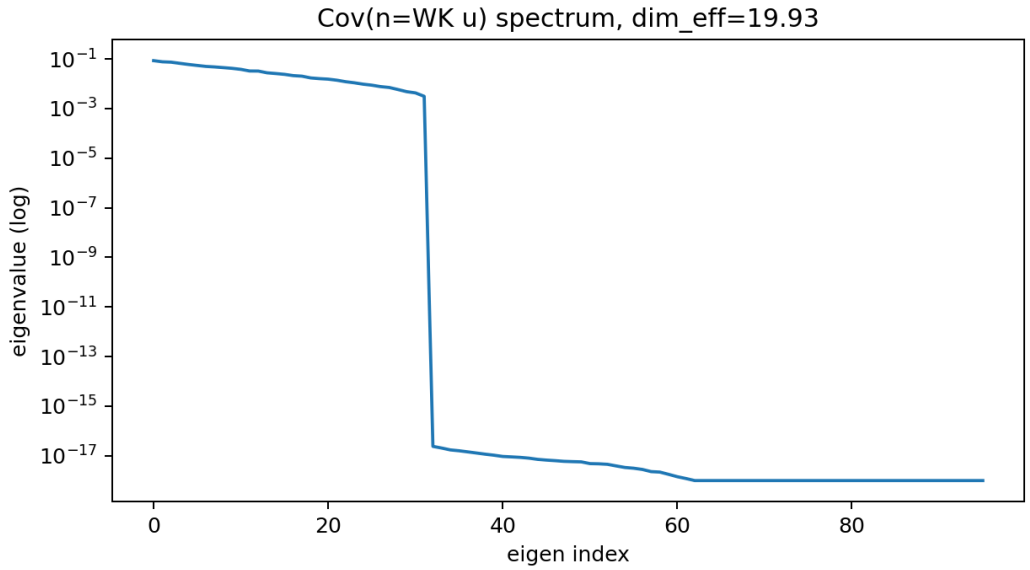

Figure 6: Spectrum of $\mathrm{Cov}(n)$ for token-space normals $n = W_K u$. The sharp drop reflects that n lives in a restricted subspace even when u has broader support.

A Appendix: Operator identities and linear changes of variables

This appendix collects exact algebraic identities used repeatedly in the main text.

A.1 Tilt pullback under linear maps

For $u \in \mathbb{S}^{d-1}$ and token vectors $x_i \in \mathbb{R}^{d_{\text{model}}}$,

$$u^\top k_i = u^\top W_K^\top x_i = (W_K u)^\top x_i. \tag{64}$$

Thus a key-space projection along u corresponds exactly to a token-space projection along normal $W_K u$. For a query x_j, the induced token-space normal is

$$n_j := W_K \frac{W_Q^\top x_j}{\|W_Q^\top x_j\|}. \tag{65}$$

This makes the per-query measurement geometry explicit.

A.2 Bilinear interaction operator

The logit operator factors through the head dimension d via $B = W_Q W_K^\top$:

$$\ell_{ji} = \frac{x_j^\top B x_i}{\sqrt{d}}, \qquad \mathrm{rank}(B) \le d. \tag{66}$$

The inequality follows from $\mathrm{rank}(W_Q W_K^\top) \le \min\{\mathrm{rank}(W_Q), \mathrm{rank}(W_K)\} \le d$.

B Appendix: Computing null spaces in practice

This appendix summarizes the three bottlenecks as concrete computations.

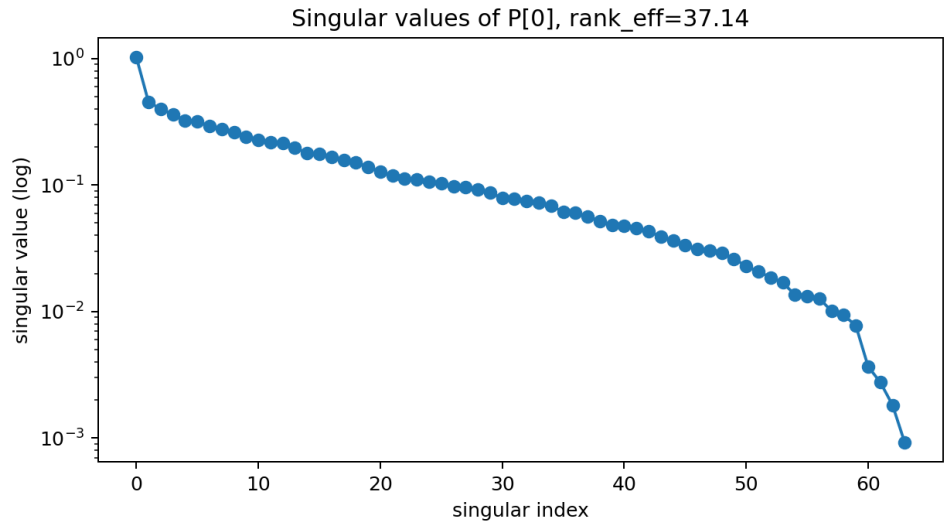

Figure 7: Singular values of a representative attention matrix P (log scale). The effective rank summarizes the spectral spread and serves as a routing-collapse diagnostic.

B.1 Hard parameter nulls

Compute $\ker(W_K^\top)$ and $\ker(W_Q^\top)$ by SVD of the parameter matrices. For the bilinear form $B = W_Q W_K^\top$, compute an SVD $B = U\Sigma V^\top$; right singular vectors corresponding to zero singular values span $\ker(B)$.

B.2 Dataset limited-angle nulls

Given a dataset, collect per-token tilts $u_j = q_j/\|q_j\|$ and compute the empirical span (e.g., via SVD of the matrix whose rows are $u_j^\top$). Directions orthogonal to this span form the limited-angle null $(\mathrm{span}\{u_j\})^\perp$.

B.3 Forward-pass routing nulls

For a forward pass, compute the attention matrix P. An SVD $P = U\Sigma R^\top$ yields $\ker(P)$ as the span of columns of R corresponding to zero singular values. The dimension satisfies $\dim\ker(P) = n_k - \mathrm{rank}(P)$.

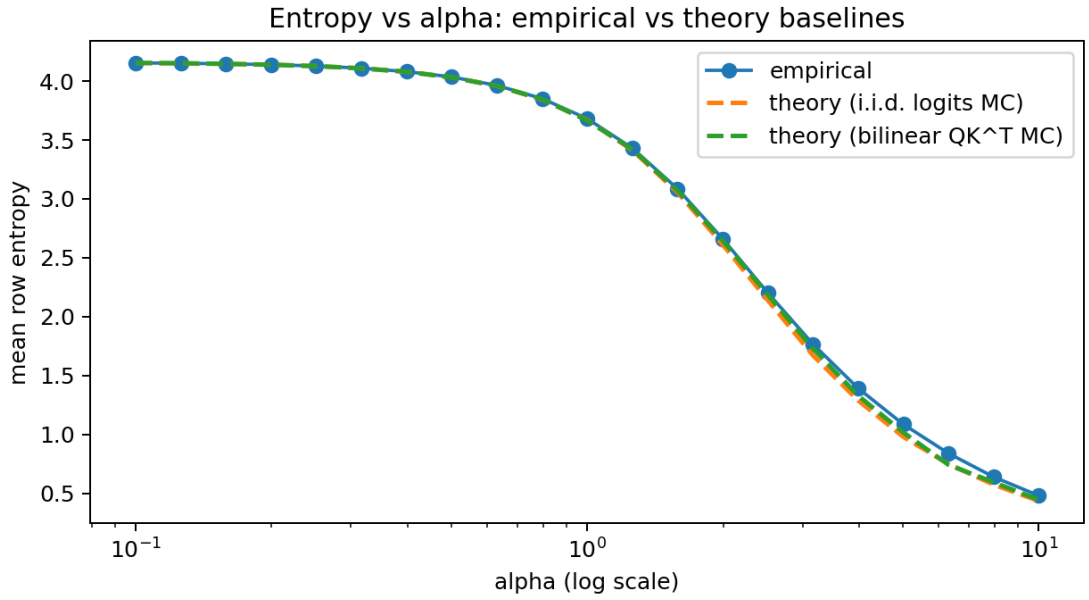

Figure 8: Mean row entropy vs logit scale α. Both mean-field baselines capture the monotone entropy drop as attention sharpens.

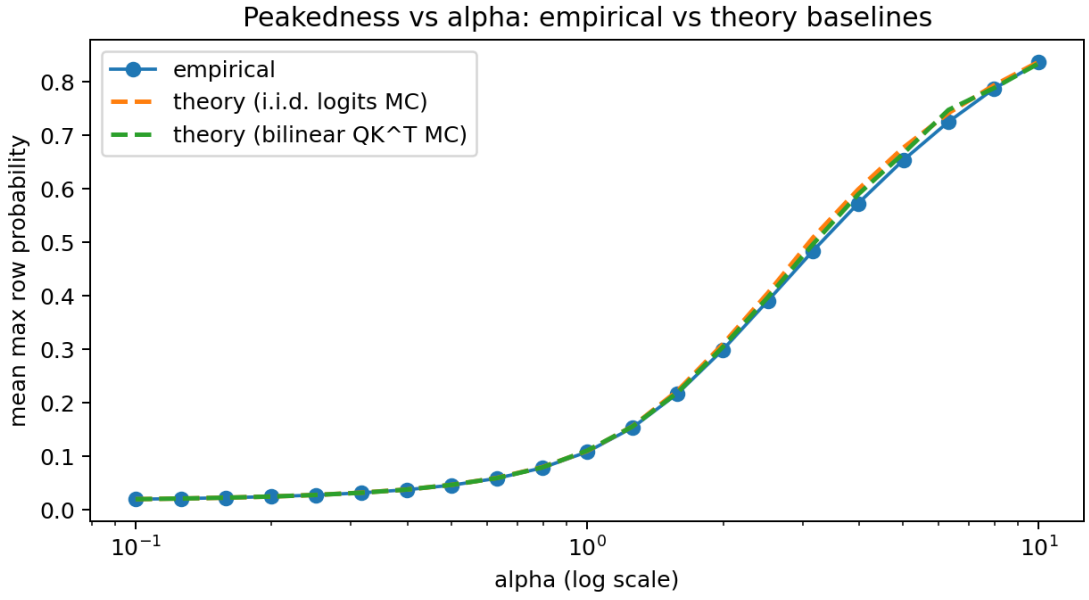

Figure 9: Mean max row probability vs α. Peakedness increases monotonically with effective temperature, and the mean-field baselines closely track the empirical curve.

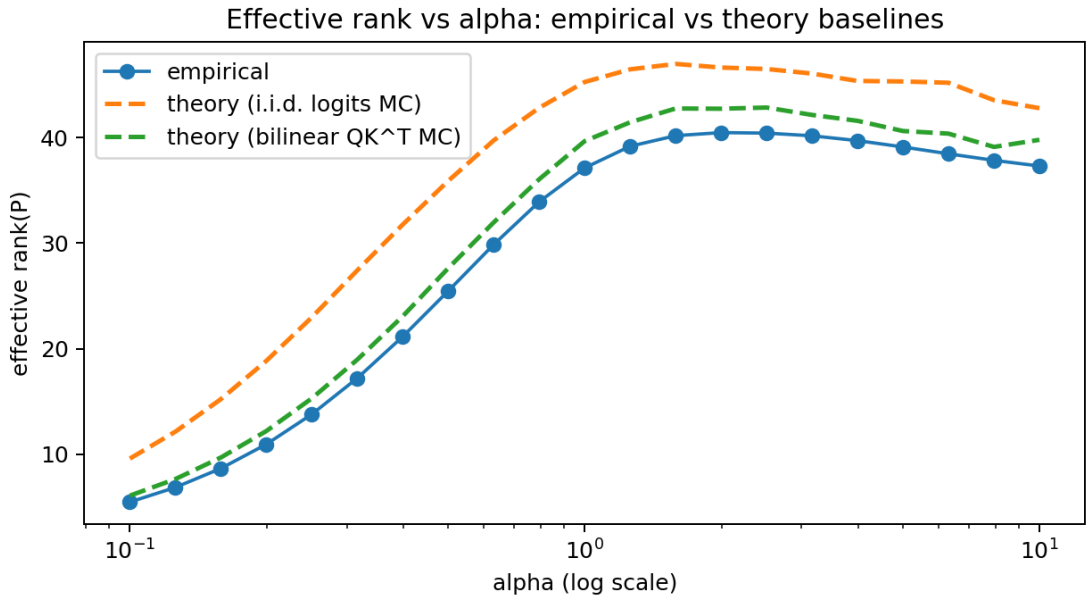

Figure 10: Effective rank(P) vs α. Rank increases rapidly from the near-uniform regime and then saturates/declines as attention becomes one-hot. The structured bilinear baseline is closer than i.i.d. logits, reflecting the importance of matrix-level correlations.