A NORMAL FORM FOR SOFTMAX ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We show that every softmax attention head admits a unique three-term normal form with interpretable components. An inspired choice of gauge in the score function clearly separates geometric and topological contributions, including a metric term which is the complement of a Laplacian in a head-induced geometry, a 1-cochain whose circulation encodes the area of token triangles, and a rank-one offset that biases the keys independently of the queries. This enables a precise characterization of when softmax attention is purely diffusive and when differently represented discrete Laplacian operators are the same. Applying the proposed normal form shows that softmax attention generically implements a quasilinear advection-diffusion equation satisfying a discrete maximum principle, guaranteeing stable information propagation regardless of its weights. Additional analysis shows that the metric and circulation terms in the normal form determine a guaranteed rate at which a softmax head mixes tokens, while the remaining offset term bounds the amount of attention that each key can receive. Ultimately, this leads to a reinterpretation of observed phenomena related to attention sinks and oversmoothing that may be useful for understanding deep attention networks.

1 INTRODUCTION

Softmax attention is the mechanism most language models use to aggregate information from across positions in an input sequence. By allowing each positional index in the sequence to attend to token information from other indices, architectures such as Transformers (Vaswani et al., 2017) can make predictions that are informed by positional context, leading to improved coherence and overall semantic utility. On the other hand, modern attention mechanisms have documented issues. Repeated aggregation operations are known to degrade representational quality, functioning as a form of Laplacian smoothing that homogenizes token information in the deep layer limit (Li et al., 2018). Particularly in graph domains (where a Transformer can be considered as operating on a fully connected graph), certain classes of graph neural networks (GNNs) (Corso et al., 2024) such as graph convolutional networks (GCNs) (Kipf & Welling, 2017) can converge exponentially to features that retain only connectivity and degree information (Oono & Suzuki, 2020), a failure mode known as *oversmoothing*.

Since softmax attention adaptively modifies the conductance (or weight) of each edge in the graph based on the token information at adjacent nodes, it is frequently claimed that attention substantially mitigates oversmoothing. However, the documented effects of attention on this failure mode is mixed. Pure self-attention loses rank doubly exponentially with depth (Dong et al., 2021), and other attention-based GNNs like graph Transformers lose expressive power exponentially (Wu et al., 2023b). Because of this, it is common to reason about attention heads by treating them as diffusion operators. Indeed, any row-stochastic matrix can be written as $A = I - L$ with $L1 = 0$, and therefore L can be interpreted as the random-walk Laplacian of a weighted graph when it is symmetric (c.f. Remark 2.2). Therefore, the "frozen-head" attention update $F \mapsto AF$ (with A constant) is then an explicit Euler step of a feature-dependent graph heat equation (Chamberlain et al., 2021a; Wu et al., 2023a; Sander et al., 2022; Gruber et al., 2023).

With this said, the diffusion interpretation of softmax attention requires restrictive symmetry assumptions that are generally unsatisfied by learned query and key weights. Previous work that explicitly adds transport or reaction terms to attention networks suggests that non-diffusive phenomena in attention networks are important (Zhao et al., 2023; Gruber et al., 2023; Eliasof et al., 2024; Choi et al., 2023), but it remains unclear what the composition of a trained softmax head

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.

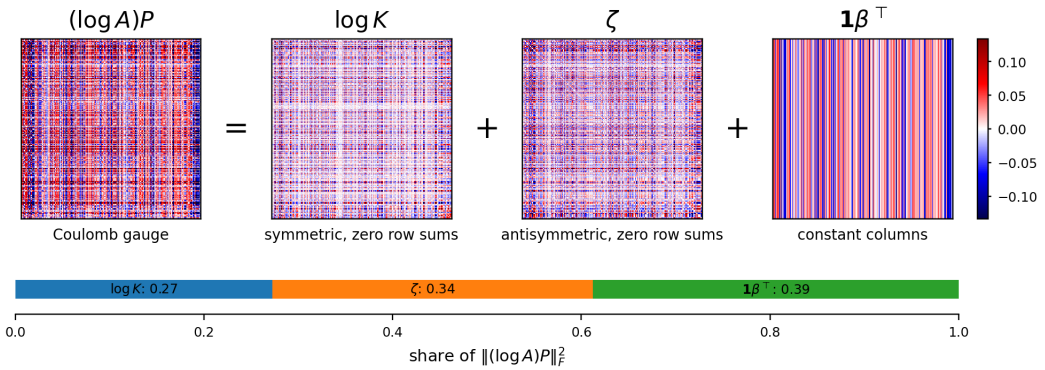

Figure 1: The normal coordinates of a softmax head, trained on Citeseer as part of a residual-connected GNN and queried on a complete graph extension of a nodal subset.

actually is or how it operates mechanistically. This paper provides an answer for a single softmax head on a connected graph: a softmax head A is the complement of a metric graph Laplacian (and therefore diffusive) precisely when the skew-symmetric part of its elementwise logarithm $\log A$ is exact as a 1-cochain (see Theorem 2.3). In fact, we further show that orthogonally decomposing the row-centered Coulomb representative of $\log A$, also known as the "energy field" in Lee (2026), yields a the unique *normal form* (see Theorem 3.4) with interpretable meaning,

$$A = \mathrm{softmax}_j(\log K + \zeta + 1\beta^{\mathsf{T}})\,.$$

In the scaled dot-product case, the symmetric $\log K$ realizes a pullback metric on sequence positions inherited from tokens, the skew-symmetric and divergence-free ζ measures token-space areas, and β is a query-independent offset vector acting as an additive key bias and related to attention sinks (Gu et al., 2025; Xiao et al., 2024; Barbero et al., 2025). As it turns out, this normal form has a variety of interesting applications; here, we use it to show that softmax attention implements an advection-diffusion equation, and that different behaviors in softmax heads are controlled by different normal form components.

To summarize, our contributions are the following: **(1)** We formulate a quantitative condition for softmax heads to be purely diffusive (see Theorem 2.3) and organize common GNN Laplacians as instances of a single construction (see Table 1); **(2)** We prove that every softmax head has a unique gauge-invariant normal form that is easily interpretable in the scaled dot-product case (see Corollary 3.6); **(3)** We show that softmax attention implements a forward Euler step of a quasilinear advection-diffusion equation (see Theorem 4.1) that is always contractive; **(4)** For fully connected graphs, we show that attention yields a guaranteed smoothing factor depending only on $Y = \log K + \zeta$, and that each entry A_{ij} is bounded in a band around $\mathrm{softmax}(\beta)_j$. This recovers the cubic smoothing rate of Dong et al. (2021) for scaled dot-product heads, and also shows that ζ is bounded by the largest token-triangle area, meaning that heads approach a metric Laplacian as depth increases.

2 The Problem Setting

Consider a connected and undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ with $V = |\mathcal{V}|$ nodes, indexed by sequence positions i, and $E = |\mathcal{E}|$ edges, indexed by pairs (i, j) with $i < j$. It will be convenient to augment the edge set with V additional self-loops (i, i), yielding $\bar{\mathcal{E}}$ with $|\bar{\mathcal{E}} = E + V$. The primary object of study is the *softmax attention* (Devlin et al., 2019; Dosovitskiy et al., 2021), defined for query/key matrices $Q, K \in \mathbb{R}^{V\times d}$ and value matrices $V \in \mathbb{R}^{V\times n}$ as

$$[A(Q,K)V]_i \;=\; \textstyle\sum_{j\sim i} \mathrm{softmax}_j\big(a(\vec{q}_i, \vec{k}_j)\big)\vec{v}_j, \tag{1}$$

where $j \sim i$ when $(i, j) \in \bar{\mathcal{E}}$, $a : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a score function, $\vec{f}_i$ denotes the ith row of the matrix F, and $[\mathrm{softmax}_j M]_{ij} = \exp M_{ij} \,/\, \sum_{k\sim i} \exp M_{ik}$ is the row-wise softmax. Here and throughout the work, the function $\exp M$ and its inverse $\log M$ act element-wise on the support of

the matrix M. We adopt the convention that indices M_{ij} and matrix expressions like $uv^\intercal \in \mathbb{R}^{V\times V}$ are defined relative to the connectivity of the graph, i.e., only at indices i, j connected by an edge.

An important example of this construction is the scaled dot-product attention (Vaswani et al., 2017),

$$A(Q,K)V \;=\; \mathrm{softmax}_j\left(\tfrac{QK^\intercal}{\sqrt{d}}\right)V, \tag{2}$$

which is a special case softmax attention on the complete graph $\mathcal{G}$ with score matrix $a_{ij} = a(\vec{q}_i, \vec{k}_j) = d^{-1/2}\langle \vec{q}_i, \vec{k}_j\rangle$ and a self-loop at every vertex. It is typical for the matrices Q, K to be given in terms of a common feature matrix $F \in \mathbb{R}^{V\times n}$ and weight matrices $W_Q, W_K \in \mathbb{R}^{n\times d}$, so that $Q = FW_Q$ and $K = FW_K$. As is customary, the rows of F will be called *tokens* while its columns will be called *features*. Letting $W = d^{-1/2}W_QW_K^\intercal$, the scaled-dot score matrix can be regarded as the quadratic object $a(F) = FWF^\intercal$ with query-dependent rows and key-dependent columns. Moreover, the relevant softmax head can be written as $A(F)_{ij} = \mathrm{softmax}_j\, a(F)_{ij}$.

Graph Geometry. It will be useful to examine softmax attention using tools from the graph calculus (Lim, 2020; Grady & Polimeni, 2010). Recall that a *k-clique* is an ordered subset of $(k+1)$ mutually adjacent nodes in the sequence space $\mathcal{V}$, and a *k-cochain* is a real-valued alternating function on k-cliques $[v_{i_0}, ..., v_{i_k}]$. A fundamental topological operator on any graph is the *coboundary* d, identified for each k with the incidence matrix of k-cliques on $(k+1)$-cliques. At lowest order, this yields the graph gradient $\mathrm{d} : \mathbb{R}^V \to \Omega^1(\mathcal{G})$ with components $\mathrm{d}f_{ij} = f_j - f_i$ on node features $f \in \mathbb{R}^V$. Here, $\Omega^1(\mathcal{G})$ denotes the space of 1-cochains on the graph $\mathcal{G}$, which are identified with $V \times V$ skew-symmetric matrices supported on edges in $\mathcal{E}$. More generally, k-cochains in $\Omega^k(\mathcal{G})$ are identified with totally antisymmetric $(k+1)$-tensors on $\mathcal{V}$.

We will call a *geometry* on $\mathcal{G}$ a specification of metric data (M, m), where $M \in \mathbb{R}^{V\times V}$ is a symmetric and entry-wise nonnegative matrix of weights called *edge conductances* and $m \in \mathbb{R}^V$ is a positive vector of node weights. This is also known as a "weighted graph" in Grigor'yan (2018); Keller et al. (2021). While the data (M, m) can be formally independent, we will say that this geometry is *compatible* whenever $m = M1$, in which case m is the *induced volume*. The specification of a geometry also endows the node and edge features with natural inner products, or discrete metric tensors, enabling length and angle measurement: for $f, g \in \mathbb{R}^V$ and $U, V \in \Omega^1(\mathcal{G})$,

$$\langle f, g\rangle_m \;=\; \textstyle\sum_{i=1}^V m_i f_i g_i \;=\; \langle f, m\odot g\rangle, \qquad \langle U, V\rangle_M \;=\; \sum_{i=1}^V \sum_{j>i} M_{ij}U_{ij}V_{ij} \;=\; \langle U, M\odot V\rangle,$$

where $\langle U, V\rangle$ denotes the half-Frobenius inner product on $\Omega^1(\mathcal{G})$.

Graph Laplacian. Many popular differential operators in machine learning and signal processing arise directly from the geometry of $\mathcal{G}$. The most important for the present purpose is the lowest-order *metric graph Laplacian* (or Laplace-Beltrami) operator $\Delta_{M,m} : \mathbb{R}^V \to \mathbb{R}^V$ that encodes the metric properties of the graph $\mathcal{G}$ (Lim, 2020; Grigor'yan, 2018). To define it, recall that any Laplacian is the divergence of a gradient, where these operators are adjoint to one another with respect to some inner product on the domain. A common choice in the graph literature is the Euclidean geometry, $(M, m) = (A_\mathcal{G}, 1)$ for $A_\mathcal{G}$ the graph adjacency, in which case the *combinatorial Laplacian* is given by $\mathrm{d}^\intercal\mathrm{d} = D_\mathcal{G} - A_\mathcal{G}$ in terms of the degree matrix $D_\mathcal{G} = \mathrm{diag}(A_\mathcal{G}1)$ and the Euclidean graph divergence $\mathrm{d}^\intercal U = -U1$ (Lim, 2020; Grady & Polimeni, 2010). Going forward, $\tilde{m} = \mathrm{diag}\, m$ will denote the diagonal matrix built from m and $L_M := \mathrm{diag}(M1) - M$ will refer to the combinatorial Laplacian with edge weights M. Proposition B.1 proved in Appendix B.1 shows that the metric Laplacian is a row-wise rescaling of the combinatorial one, leading to a useful corollary characterizing when the metric Laplacian is related to an "attention like" object.

Corollary 2.1. *The metric graph Laplacian satisfies $\Delta_{M,m} = I - \tilde{m}^{-1}M := I - A$ if and only if the geometry (M, m) is compatible, in which case $A = \tilde{m}^{-1}M$ is entry-wise nonnegative, row stochastic, and m-self-adjoint.*

Remark 2.2. The matrix A in Corollary 2.1 admits a probabilistic interpretation as a Markov transition matrix (Wang et al., 2022; Erel et al., 2025), where the entry $A_{ij} = \mathbb{P}(i \to j)$ is the probability that a random walk starting at token i moves to token j along edge (i, j). In this interpretation, $\Delta_{M,m} = I - A$ is just the random walk Laplacian (von Luxburg, 2007; Levin & Peres, 2017).

Remarkably, $\Delta_{M,m}$ unifies many constructions from the GNN literature. Table 1 contains common operators as instances of the metric graph Laplacian and represents a first contribution of this work.

In particular, Proposition B.3 in Appendix B.1 shows that the change of nodal variables $f \mapsto s \odot f$ for $s \in \mathbb{R}_+^V$ induces a spectrum-preserving change $\Delta_{M,m} \mapsto \tilde{s}\Delta_{M,m}\tilde{s}^{-1}$ in the metric graph Laplacian, enabling indexing based on the coordinates (M, m, s). This yields a simple way to see when two Laplacians correspond to the same geometry. For example, rows 2-4 in Table 1 are the same operator in different units, while the GCN Laplacian of row 5 is genuinely different due to the "renormalization trick" that changes the nodal volume from d to $d + 1$ (Kipf & Welling, 2017).

name	expression	conductance M	volume m	coordinate s
Combinatorial Laplacian	$D - A$	A	1	1
Random walk normalized Laplacian	$I - D^{-1}A$	A	d	1
Symmetric normalized Laplacian	$I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$	A	d	$\sqrt{d}$
Transposed random walk normalized Laplacian	$I - AD^{-1}$	A	d	d
GCN Laplacian	$I-(D+I)^{-\frac{1}{2}}(A+I)(D+I)^{-\frac{1}{2}}$	$A + I$	$d+1$	$\sqrt{d+1}$
Diffusion maps α	$I - (K^{(\alpha)}1)^{-1}K^{(\alpha)}$	$K^{(\alpha)}$	$K^{(\alpha)}1$	1
Softmax attention with $\zeta = 0$	$I - A$	$K \odot bb^\intercal$	$M1$	1

Table 1: Conductance, volume, and coordinate are independent choices leading to various Laplacians. Rows sharing (M, m) share a spectrum. The final row is explained by Corollary 3.5, and $K^{(a)}$ denotes the anisotropic normalization of the kernel $K(x, y)$ (Coifman & Lafon, 2006).

Softmax Heads as Diffusion. The matrix $A = \tilde{m}^{-1}M$ from Corollary 2.1 is the row-normalization of a symmetric, entrywise nonnegative matrix on edges. In particular, *it is a softmax* when $M = \exp C$ for some symmetric score matrix $C \in \mathbb{R}^{V\times V}$. This recovers the known interpretation of a softmax attention head as the complement of a metric Laplacian when certain symmetry conditions are met or enforced. For example, Sander et al. (2022) take the weights $W_QW_K^\intercal$ as symmetric, while Gruber et al. (2023) explicitly symmetrize with $\operatorname{sym}(W_QW_K^\intercal) = \frac{1}{2}(W_QW_K^\intercal + W_KW_Q^\intercal)$. Similarly, kernel-based approaches like Geshkovski et al. (2025) assume a symmetric kernel, and energy/diffusion-based methods like Wu et al. (2023a) implicitly rely on weight symmetries. Note that these are generally restrictive assumptions; Proposition B.2 in Appendix B.1 shows that every symmetric score matrix corresponds to a softmax head $A = I - \Delta_{M,m}$, but not every head arises this way. We now show that this difference can be precisely quantified via the *balance cochain* $\theta(L) = \operatorname{skew}\log L = \frac{1}{2}\log(L \oslash L^\intercal) \in \Omega^1(\mathcal{G})$, valid for any $L \in \mathbb{R}^{V\times V}$ with symmetric support and single-signed off-diagonal entries, where $\operatorname{skew} M = \frac{1}{2}(M - M^\intercal)$.

Theorem 2.3. *Let $L \in \mathbb{R}^{V\times V}$ have connected symmetric support with $L1 = 0$ and negative off-diagonal entries. Then, $L = \Delta_{M,m}$ is a metric Laplacian if and only if the balance cochain $\theta(L)$ is exact, in which case $\theta(L) = \mathrm{d}\log\sqrt{m}$. In particular, a softmax head $A = I - \Delta_{M,m}$ is the complement of a compatible metric Laplacian if and only if $\theta(A) = \operatorname{skew}\log A$ is exact.*

Remark 2.4. Exactness of $\theta(L)$, which means that $\theta(L) \in \operatorname{span}\mathrm{d}$, is known as "Kolmogorov's criterion" in the probability literature (Kolmogoroff, 1936). Its name comes from equivalence to the condition $(\tilde{m}L)^\intercal = L^\intercal\tilde{m} = \tilde{m}L$ known as "detailed balance" (Kelly, 1979; Levin & Peres, 2017).

It is worth mentioning an immediate corollary that holds when $\mathcal{G}$ is complete, as in the case of a fully connected Transformer. The exactness condition on θ can be converted to a closedness condition $\mathrm{d}\theta_{ijk} = \theta_{ij} + \theta_{jk} + \theta_{ki} = 0$ on the graph curl $\mathrm{d}\theta$ (Lim, 2020).

Corollary 2.5. *When $\mathcal{G}$ is complete, a softmax head $A \in \mathcal{A}$ is the complement of a compatible metric Laplacian if and only if $\mathrm{d}\theta(A) = \mathrm{d}\operatorname{skew}\log A = 0$, i.e., its balance cochain is closed.*

3 THE NORMAL FORM OF SOFTMAX ATTENTION

By now, it is apparent that softmax heads are not automatically metric operators, but their deviation from one is encoded in the balance cochain θ. Proposition B.2 and Theorem 2.3 together yield some additional insight, but do not provide a unique decomposition of softmax attention in terms of metric and non-metric parts. The goal of this section is to exploit the gauge freedom in the softmax to expose a more nuanced *normal form* for softmax attention heads. This will enable a unique parameterization of softmax heads with interpretable terms of both geometric and topological significance.

Gauge Invariance and Invariants. A primary conclusion of this work is that the behavior of softmax attention is largely dominated by its score function, which maps nonuniquely onto softmax attention heads. More precisely, let $\mathcal{A} = \{A \in \mathbb{R}^{V\times V} : A_{ij} > 0, A1 = 1\}$ denote the $2E$-dimensional space of heads on $\mathcal{G}$, and recall that the softmax is the composition $\mathrm{softmax}_j = \mathrm{norm}_j \circ \exp$ of a component-wise exponential with a row-wise normalization. It is well known that this induces gauge freedom in the score matrix (Goodfellow et al., 2016; Namazifar et al., 2023; Lee, 2026): the row-wise transformation $a_{ij} \mapsto a_{ij} + g_i$ leaves the softmax invariant for any vector $g \in \mathbb{R}^V$. A consequence is that the usual decomposition $a = \mathrm{sym}\, a + \mathrm{skew}\, a$ of the score matrix into symmetric and skew parts is gauge-invariant, although invariant information can still be extracted from it. Discussing this will use the node-to-edge averaging operator $g \mapsto \mathrm{av}\, g = \mathrm{sym}(g1^\intercal)$.

Proposition 3.1. *Let $g \in \mathbb{R}^V$. Two attention scores induce the same head A if and only if they differ by $g1^\intercal$ on their support. Under $a \mapsto a + g1^\intercal$, the symmetric and skew-symmetric parts of the score satisfy*

$$\mathrm{sym}\, a \mapsto \mathrm{sym}\, a + \mathrm{av}\, g, \qquad \mathrm{skew}\, a \mapsto \mathrm{skew}\, a - \tfrac{1}{2}\mathrm{d}g.$$

In particular, the quantities $G_{ij} := \mathrm{sym}(a)_{ii} + \mathrm{sym}(a)_{jj} - 2\,\mathrm{sym}(a)_{ij}$ and $\xi = \mathrm{d}\,\mathrm{skew}\, a$ are gauge-invariant and computable from cross ratios without access to the weights of A,

$$G_{ij} = \log \tfrac{A_{ii}A_{jj}}{A_{ij}A_{ji}}, \qquad \xi_{ijk} = \tfrac{1}{2}\log \tfrac{A_{ij}A_{jk}A_{ki}}{A_{ji}A_{kj}A_{ik}}.$$

Moreover, $\xi = \mathrm{d}\theta(A)$ is also computable from the balance cochain.

Proposition 3.1 shows that the score matrix a contains gauge-invariant quantities G and ξ formed by polarizing the symmetric part and differentiating the skew part. To see how these capture salient features of the head $A = \mathrm{softmax}_j(a)$, recall that the scaled dot product attention (2) has score matrix $a(F) = FWF^\intercal$, and let $S = \mathrm{sym}\, W$ and $B = \mathrm{skew}\, W$. By convention, the coboundary d acts row-wise on matrices $F \in \mathbb{R}^{V\times n}$, so that $\mathrm{d}F_{ij} = \vec{f}_j - \vec{f}_i \in \Omega^1(\mathcal{G}) \otimes \mathbb{R}^n$ is a 1-cochain on V with values in $\mathbb{R}^n$. Considering $F : \mathcal{V} \to \mathbb{R}^n$ as a mapping between sequence space and token space yields *pullbacks* of the weight matrix components S, B by F,

$$F^*S_{ij} = \mathrm{d}F_{ij}^\intercal S\, \mathrm{d}F_{ij} = \left|\vec{f}_i - \vec{f}_j\right|_S^2, \qquad F^*B_{ijk} = \mathrm{d}F_{ki}^\intercal B\, \mathrm{d}F_{kj} = (\vec{f}_i - \vec{f}_k)^\intercal B(\vec{f}_j - \vec{f}_k),$$

where $|\vec{f}|_S^2 = \vec{f}^\intercal S\vec{f}$ is the seminorm of $\vec{f}$ defined by S and $F^*B \in \Omega^2(\mathcal{G})$ is a 2-cochain on index triples $[i,j,k]$. The following corollary of Proposition 3.1 computes the invariants (G, ξ) for a scaled dot-product head.

Corollary 3.2. *Let $a(F) = FWF^\intercal = F(S+B)F^\intercal$ be a scaled-dot score matrix. Then, $G = F^*S$ is a (potentially degenerate) pullback metric and $\xi = \mathrm{d}\theta(A) = F^*B$ is an area cochain.*

Corollary 3.2 already provides substantial geometric understanding of scaled dot product attention heads. In particular, when $S \succ 0$ is positive definite, then $G = F^*S$ is a pullback metric inherited from the tokens, and ξ measures the B-weighted area of token triples. More precisely, if $T = [\vec{f}_i, \vec{f}_j, \vec{f}_k]$ is a token triangle, then the 2-cochain value $F^*B_{ijk} = 2\int_T \omega$ is the integral of the constant differential two-form $\omega(\vec{u}, \vec{v}) := \vec{u}^\intercal B\vec{v}$ for $\vec{u}, \vec{v} \in \mathbb{R}^n$. Since ξ also encodes the balance cochain from Theorem 2.3, the following result is immediate from Corollary 2.5.

Corollary 3.3. *A scaled dot product attention head on a complete graph $\mathcal{G}$ is the complement of a metric Laplacian if and only if the area form $F^*B = 0$ vanishes, i.e, $\vec{u}^\intercal B\vec{v} = 0$ for all $\vec{u}, \vec{v} \in \mathrm{span}\, \mathrm{d}F$. Letting $\mathrm{rank}\, B = 2r$, this is impossible when $\dim \mathrm{span}\, \mathrm{d}F > n - r$.*

Corollary 3.3 is remarkable because it shows that the likelihood of A being purely diffusive declines rapidly as the number of tokens V grows. In particular, any sufficiently large Transformer whose

number of input tokens exceeds its embedding dimension is highly unlikely to be the complement of a metric Laplacian. Notice also that $F^*B = 0$ is a weaker condition on the head A than is typically assumed in the literature. In contrast to the global symmetry assumptions mentioned in section 2 that enforce $B = 0$, $F^*B = 0$ is a pullback condition and therefore fully determined by the restriction of B to $\operatorname{span} \mathrm{d}F$. Since Corollary 3.3 is both necessary and sufficient, so it is the weakest statement that ensures a scaled-dot attention head admits a diffusion interpretation.

A Unique Parameterization. We have seen that the invariants (G, ξ) from Proposition 3.1 are useful for understanding the behavior of softmax attention, but they cannot uniquely describe an attention head. To see this, observe that G carries E degrees of freedom as a symmetric and diagonal-free matrix. Similarly, ξ is totally skew-symmetric and exact, carrying $E - (V - 1)$ degrees of freedom (up to the harmonic elements of $\Omega^1(\mathcal{G})$). Therefore, the pair (G, ξ) has dimension $2E - (V - 1)$, meaning that $V - 1$ degrees of freedom remain unaccounted for in a generic softmax head $A \in \mathcal{A}$. Repairing this will yield an interesting coordinate system on softmax heads that exposes another gauge-invariant object functioning as a query-independent key bias.

Since a softmax head only determines an equivalence class of score matrices modulo $g \in \mathbb{R}^V$, establishing a unique parameterization requires gauge-fixing. A convenient choice arises from the requirement that $g = \arg\min_h |a + h1^\intercal|_F^2 = -a1 \oslash d$ produce the representative with minimal Frobenius norm (where d is the vector of nodal degrees). This is known as choosing the *Coulomb gauge* in the physics literature (Jackson, 1999; Frankel, 2012), and it ensures that representative scores are row-centered. In particular, letting $[a] = \{a + g1^\intercal : g \in \mathbb{R}^V\}$ be an equivalence class of scores and $P(a) = a - (a1 \oslash d)1^\intercal$ be Frobenius-orthogonal projection onto the complement of the gauge space, the Coulomb gauge is the mapping $[a] \mapsto P(a)$ where $P(a)1 = 0$. When $\mathcal{G}$ is complete, this reduces to $P(a) = aP$ for $P = I - \frac{1}{V}11^\intercal$. It follows that the softmax is a bijection from the set of row-centered score matrices onto attention heads $\mathcal{A} \in \mathcal{A}$ with inverse $A \mapsto P(\log A)$. Decomposing these matrices yields an interesting *normal form* for softmax attention.

Theorem 3.4. *Every attention head $A \in \mathcal{A}$ has a unique normal form*

$$A = \operatorname{softmax}_j(\log K + \zeta + 1\beta^\intercal), \qquad A_{ij} = \frac{K_{ij} e^{\beta_j} e^{\zeta_{ij}}}{\sum_{k \sim i} K_{ik} e^{\beta_k} e^{\zeta_{ik}}},$$

where the normal coordinates $(\log K, \zeta, \beta)$ are Frobenius-orthogonal, diffeomorphic onto $\mathcal{A}$, and

$$(\log K)^\intercal = \log K, \quad (\log K)1 = 0, \qquad \zeta^\intercal = -\zeta, \quad \mathrm{d}^\intercal \zeta = -\zeta 1 = 0, \qquad 1^\intercal \beta = 0.$$

Moreover, these coordinates satisfy $\dim \log K = E$, $\dim \zeta = E - V + 1$, and $\dim \beta = V - 1$.

Observe that the normal coordinates $(\log K, \zeta, \beta)$ separate the head A into non-redundant orthogonal components. Moreover, computing them is straightforward: simply project $\log A$ onto the complement of the space $\mathcal{B} = \{u1^\intercal + 1v^\intercal : u, v \in \mathbb{R}^V\}$ and then symmetrize (see Appendix B.2). Note that the symmetry of the first two coordinates $(\log K, \zeta)$ combined with the centering conditions $(\log K)1 = \zeta 1 = 0$ ensure that all key-only dependence has been collected into the third coordinate β. In view of Theorem 3.4, we will write $A \simeq (\log K, \zeta, \beta)$ to describe the normal-form-represented softmax head described by a given set of normal coordinates, which is unique up to the gauge action. To further understand the distinct meaning of each term in the normal form, it is useful to collect an immediate corollary that determines when a softmax head prescribed by normal coordinates is the complement of a metric graph Laplacian.

Corollary 3.5. *Let $A \simeq (\log K, \zeta, \beta)$ be a softmax head and denote $b = e^\beta$. Then, A has the expression*

$$A \;=\; \operatorname{norm}_j(K \odot 1b^\intercal \odot e^\zeta) \;=\; \operatorname{norm}_j(M \odot e^\zeta), \qquad M = K \odot bb^\intercal.$$

In particular, $A = I - \Delta_{M,m}$ is the complement of a metric Laplacian if and only if $\zeta = 0$.

Corollary 3.5 connects the normal form transparently to the metric Laplacian from Proposition B.1, and also connects the first and second normal coordinates $(\log K, \zeta)$ to the gauge invariants (G, ξ) from Proposition 3.1. Indeed, it follows that $\theta(A) = \zeta + \frac{1}{2}\mathrm{d}(\beta - g)$ for some $g \in \mathbb{R}^V$, so that $\theta(A)$ being exact implies that the circulation ζ is exact and divergence-free, hence $\zeta = 0$ captures topological information. Moreover, the metric invariant $G_{ij} = \log K_{ii} + \log K_{jj} - 2 \log K_{ij}$ is exactly recoverable from $\log K$ alone. What remains in the normal form is a rank-one offset $1\beta^\intercal$ supported on $\mathcal{G}$ that is neither symmetric nor skew-symmetric. This leftover term can be related to

attention sinks (Xiao et al., 2024; Gu et al., 2025; Barbero et al., 2025) in the Transformer case. That is, β is the query-independent part of the head that consumes a large fraction of attention: if $\bar{q} = \frac{1}{V} Q^\intercal 1$ and $\bar{k} = \frac{1}{V} K^\intercal 1$ denote query/key means, then $\beta_j = d^{-1/2}\langle \bar{q}, \vec{k}_j - \bar{k}\rangle$ measures the alignment of the jth key with the mean query, meaning that β functions like an additive key bias (see Gu et al. (2025)). Theorem 4.3 will show that β_j also controls the attention value A_{ij}. As before, it is informative to compute the normal form for the scaled dot product attention.

Corollary 3.6. *Let $a(F) = FWF^\intercal = F(S+B)F^\intercal$ be the scaled-dot score matrix on a complete graph $\mathcal{G}$. Then, the normal coordinates $(\log K, \zeta, \beta)$ are given by*

$$\log K = HSH^\intercal = -\tfrac{1}{2} P(F^* S)P, \qquad \zeta = HBH^\intercal = \tfrac{1}{V}\, \mathrm{d}^\intercal F^* B, \qquad \beta = HW^\intercal \bar{f},$$

where $H = PF$ are the token-centered features and $\bar{f} = \frac{1}{V} F^\intercal 1$ is the token centroid. In particular, $\log K$ and ζ are invariant under translations of F, $\zeta = 0$ if and only if the pullback $F^* B = 0$ vanishes, and $\beta = 0$ exactly when $W^\intercal \bar{f} \in \ker H$.

Corollary 3.6 confirms what was seen in the general case: the pullback metric $F^* S$ is encoded in $\log K$, the circulation ζ is a primitive for the area form $F^* B$, and the attention sink β measures how much each centered token aligns with the token centroid $\bar{f}$. This also supports the observation of massive activations in language models (Sun et al., 2024), which enlarge the norm $|\bar{f}|$ of the centroid and hence increase the value of β. Notice that common scaling operations like LayerNorm (Ba et al., 2016) that shift the features but not the tokens do not remedy this, but token-centering does.

4 Applications of the Normal Form

The normal form in Theorem 3.4 (and Corollary 3.5) provides a complete description of any softmax attention head A that can be leveraged for various applications. This section will discuss two related to depth-wise dynamics. Commonly, the inference-time behavior of attention heads is explained in terms of spectral properties of the attention graph, which implicitly relies on a metric Laplacian operator, and sometimes the interpretation of attention as adaptive diffusion (Chamberlain et al., 2021a; Topping et al., 2022; Di Giovanni et al., 2023). We first refine this understanding by studying the propagation rule $F^{(n+1)} = A(F^{(n)})F^{(n)}$ in the context of the normal form. This can be considered as a forward Euler step of the equation $\dot{F} = -(I - A)F$, which is equivalent to the gradient flow $\dot{F} = -\Delta_{M,m} F$ when (and only when) the softmax head A is the complement of a metric Laplacian.

Softmax Advection-Diffusion. The first application of Corollary 3.5 will show that the continuous update $\dot{F} = -(I - A(F))F$ takes the form of a quasilinear advection-diffusion equation, which relies on the adjoint $\mathrm{av}^*_{M,m}$ of the av operator discussed in Appendix B.3. This provides a refinement of treatments like Chamberlain et al. (2021a;b) that treat softmax attention as pure diffusion.

Theorem 4.1. *Consider a softmax head $A \simeq (\log K, \zeta, \beta)$ and let $M = K \odot bb^\intercal$ with $b = e^\beta$. Consider the geometry $(\bar{M}, m) = \big(M \odot \cosh\zeta, (M \odot e^\zeta)1\big)$. Then, it follows that*

$$\dot{F} \;=\; (A - I)F \;=\; -\Delta_{\bar{M},m} F + \mathrm{av}^*_{\bar{M},m}(u \odot \mathrm{d}F),$$

where $u = 2\tanh\zeta \in \Omega^1(\mathcal{G})$ is the velocity cochain. Moreover, $\bar{M}1 = m \odot \big(1 + \frac{1}{2}\mathrm{d}^*_{\bar{M},m} u\big)$ and $A^\intercal m = m \odot (1 + \mathrm{d}^*_{\bar{M},m} u)$, so that $(\bar{M}, m)$ is compatible and $A^\intercal m = m$ if and only if u is divergence-free.

Theorem 4.1 is analogous to the continuous PDE $\dot{F} = \mathrm{div}(\bar{M}(F)\nabla F) - u \cdot \nabla F$, where the advection term $u \cdot \nabla F = \mathrm{div}(uF) - (\mathrm{div}\, u)F$ is L^2-skew-symmetric if and only if $\mathrm{div}\, u = 0$ and the flow is incompressible. Here, this condition is replaced with incompressibility of the velocity cochain, i.e. $\mathrm{d}^*_{\bar{M},m} u = 0$, which enables advection but not compression; an incompressible attention flow satisfies $A^\intercal m = m$ and has zero flux of token mass at each node. In contrast to the general case, it is also remarkable that the attention flow in Theorem 4.1 satisfies the velocity bound $|u| \leq 2$ independent of the normal form circulation ζ. This means that large circulation values ζ_{ij} cannot accelerate edge-wise transport but do produce edges with high conductance $\bar{M}$. The following result discusses consequences of this on the *diameter* of the token cloud $\delta(F) = \max_{i,j} |\vec{f}_i - \vec{f}_j|$.

Corollary 4.2. *Let $A \simeq (\log K, \zeta, \beta)$ and let $(\bar{M}, m)$ and u be as in Theorem 4.1. Then, the solution $F(t)$ to $\dot{F} = (A - I)F$ satisfies the discrete maximum principle $\delta(F(t)) \leq \delta(F_0)$ for all initial F_0.*

This result shows that generic (softmax) attention dynamics are always diffusion-dominated despite not being a metric heat flow, explaining why the study of attention as a diffusion process has been so successful. Since advection never outweighs diffusion on any edge in $\mathcal{G}$, the forward-Euler discretization $F^{(\ell+1)} = A(F^{(\ell)})F^{(\ell)}$ is stable, and each application of the head A shrinks the diameter of the token cloud, eventually homogenizing all information contained in F. Corollary B.5 in Appendix B.3 specializes Theorem 4.1 to the scaled dot-product case.

Oversmoothing and Oversquashing. The next application of the normal form involves oversmoothing and oversquashing phenomena in deep attention networks. Recall that oversmoothing is a failure mode commonly quantified via the *Dirichlet energy* (see Appendix A), defined for edge conductance M and constant SPD target metric N as $\mathcal{D}_{M,N}(F) = \frac{1}{2}\operatorname{tr}_M(F^*N)$. While exponential decrease in the Dirichlet energy is a good proxy for smoothing on static graphs where the conductance M is constant, this is not the case for attention graphs where $M = M(F)$ depends on the evolving features. In fact, $\mathcal{D}_{M,N}$ may actually increase despite the fact that token representations are converging (c.f. Appendix A and Appendix C.3). Instead, we will say a network *oversmooths* when the radius $\delta(F^{(\ell)}) \to 0$ exponentially as $\ell \to \infty$, so that the diameter of the token cloud shrinks to zero with increasing depth. To discuss this, it is useful to define the *spread* $\kappa(A) := \max_i \operatorname{osc}_{j\sim i} A_{ij}$ in terms of the oscillation $\operatorname{osc}_{j\sim i} A_{ij} = \max_{j\sim i} A_{ij} - \min_{j\sim i} A_{ij}$. Letting $A \simeq (\log K, \zeta, \beta)$ be an attention head and denoting $Y = \log K + \zeta$ along with $\hat{b} = \operatorname{softmax}\beta$, the following result shows that the minimal rate of smoothing is controlled by the spread of Y while the attention values are controlled by β.

Theorem 4.3. *Let $\mathcal{G}$ be complete and $A \simeq (\log K, \zeta, \beta)$. Then, the diameter $\delta(AF) \le (1 - e^{-\kappa(Y)})\delta(F)$ decreases monotonically under repeated applications of A at a guaranteed rate depending on Y. Moreover, the amount of attention that token i pays to token j is bounded by $e^{-\kappa(Y)}\hat{b}_j \le A_{ij} \le e^{\kappa(Y)}\hat{b}_j$, and $A = 1\hat{b}^\intercal$ exactly when $Y = 0$.*

Theorem 4.3 shows that there is a guaranteed rate of token smoothing under the softmax head A that depends only on the first two normal coordinates $(\log K, \zeta)$. Conversely, the final coordinate β bounds the amount of attention paid to each key, with its softmax forming the center of a band modulated by Y. This is the complete graph interpretation of squashing: a key j is unimportant to query i roughly when it is unimportant to every query. It is useful to collect a further consequence of Theorem 4.3 on the scaled-dot case that holds for GNNs with and without residual connection,

$$F^{(\ell+1)} = F^{(\ell)} + \sigma\big(A(F^{(\ell)})F^{(\ell)}W_V\big), \qquad F^{(m+1)} = \sigma\big(A(F^{(m)})F^{(m)}W_V\big),$$

where the activation function σ is assumed to be L_σ-Lipschitz. To state it, let $\operatorname{area}(F) = \max_{i,j,k} \operatorname{area}(\vec{f}_i, \vec{f}_j, \vec{f}_k)$ be the largest Euclidean area of a token triangle, and let $\delta^{(m)} = \delta(F^{(m)})$.

Corollary 4.4. *Let $a(F) = FWF^\intercal$ be a scaled dot-product score and $A = \operatorname{softmax}_j a$. Then, $\kappa(Y) \le \|W\|\delta(F)^2$ and $|\zeta_{ij}| \le 2\|B\| \operatorname{area}(F)$. In particular, for the networks $F^{(\ell)}$ and $F^{(m)}$ defined above with $\rho = L_\sigma\|W_V\|$, it follows that*

$$\delta^{(m+1)} \le \rho\big(1 - e^{-\|W\|(\delta^{(m)})^2}\big)\delta^{(m)}, \qquad |\delta^{(\ell+1)} - \delta^{(\ell)}| \le \rho\big(1 - e^{-\|W\|(\delta^{(\ell)})^2}\big)\delta^{(\ell)}.$$

In particular, $\delta^{(m+1)} \le c(\delta^{(m)})^3$ without a residual connection, so that $\delta^{(m)}$ collapses doubly exponentially when $c(\delta^{(m)})^2 < 1$. On the other hand, $\delta^{(\ell+1)} \ge (1 - c(\delta^{(\ell)})^2)\delta^{(\ell)}$ with a residual connection, so that each layer changes $\delta(F^{(\ell)})$ by a factor of $1 + \mathcal{O}((\delta^{\ell})^2)$.

Corollary 4.4 shows that oversmoothing in scaled dot-product attention is self-accelerating and recovers the cubic rate of Dong et al. (2021) through a different means. At the same time, it also confirms that attention heads become more Laplacian-like with token collapse, since ζ decays with the maximum area of token triples and (by Corollary 3.6) vanishes exactly when $F^*B = 0$. As expected, adding a residual connection substantially mitigates this effect.

5 Experiments

Here we present some experiments on trained language models that illustrate the theoretical contributions. Additional experiments are contained in Appendix C

Energy share in pre-trained encoders. To test the proportionality of normal-form components, evaluate BERT-base and BERT-large (Devlin et al., 2019), RoBERTa-base (Liu et al., 2019),

and DistilBERT (Sanh et al., 2019) on 96 deterministically selected, 128-token blocks from the WikiText-2 test split (Merity et al., 2017). For every layer and head, we reconstruct the scaled dot-product score $a = d^{-1/2}QK^\intercal$ and compute $(\log A)P = aP = \log K + \zeta + 1\beta^\intercal$ for $P = I - V^{-1}11^\intercal$, reporting the relative squared Frobenius energy of each orthogonal coordinate. We also isolate the query-dependent geometry from the independent sink coordinate using $r_{\mathrm{met}} = |\log K|_F^2/(|\log K|_F^2 + |\zeta|_F^2)$. The results are displayed in Table 2, using a randomly initialized BERT-base for comparison (see also Figure 2). Observe that the $\log K$ share exceeds that of the baseline in every case, and that the $1\beta^\intercal$ share grows with depth. Importantly, the ζ share (computed as one minus the others) is nonzero in every model, confirming that trained softmax heads are not generically diffusive. The increase in r_{met} indicates increased diffusivity with depth, consistent with Corollary 4.4.

Table 2: Normal-form energy across pre-trained encoders. Entries are averages ± 95% confidence over heads and selected layers. Early uses $1, ..., \lceil L/4 \rceil$ and late uses $\lfloor 3L/4 \rfloor + 1, ..., L$.

	$\log K$ share		$1\beta^\intercal$ share		r_{met}	
Model	Early	Late	Early	Late	Early	Late
BERT-base	0.596 ± 0.002	0.471 ± 0.007	0.251 ± 0.003	0.452 ± 0.006	0.786 ± 0.001	0.853 ± 0.004
BERT-large	0.321 ± 0.002	0.425 ± 0.006	0.554 ± 0.003	0.488 ± 0.006	0.683 ± 0.001	0.830 ± 0.004
RoBERTa-base	0.626 ± 0.002	0.628 ± 0.004	0.171 ± 0.002	0.308 ± 0.003	0.757 ± 0.001	0.903 ± 0.002
DistilBERT	0.627 ± 0.002	0.545 ± 0.005	0.208 ± 0.002	0.392 ± 0.004	0.787 ± 0.001	0.892 ± 0.003
BERT-base (random)	0.320 ± 0.001	0.212 ± 0.001	0.367 ± 0.001	0.582 ± 0.001	0.506 ± 0.000	0.507 ± 0.000

Intervening on Normal Coordinates. The next experiment investigates the performance of trained softmax heads when different components of the normal form are excluded post-training. Table 3 shows the averaged result of zeroing the $\log K$ and ζ components of each head in the final depth quartile for the same language models (see also Figure 3). Observe that ζ holds only 6-9% of late-layer energy, but removing it changes outputs by as much as 47%. While not tabulated, it is remarkable that setting $\beta = 0$ changes the head output substantially, producing relative errors of 150-170% and supporting empirical observations regarding attention sinks. As expected, query-centering removes the measured β energy fraction completely, whereas LayerNorm leaves fractions between 0.321 and 0.469.

Table 3: Ablation study on normal coordinates. Errors are sequence means ± 95% confidence interval; "cosine" indicates cosine similarity.

Model	$\zeta = 0$ error	$\zeta = 0$ cosine	$\log K = 0$ error	only $1\beta^\intercal$ error
BERT-base	0.406 ± 0.003	0.910 ± 0.002	0.780 ± 0.006	0.853 ± 0.005
BERT-large	0.441 ± 0.003	0.888 ± 0.002	0.839 ± 0.003	0.900 ± 0.002
RoBERTa-base	0.471 ± 0.003	0.865 ± 0.002	0.793 ± 0.003	0.816 ± 0.003
DistilBERT	0.351 ± 0.002	0.932 ± 0.001	0.745 ± 0.005	0.811 ± 0.004

6 Conclusion

We showed that every softmax attention head has a unique normal form with interpretable and Frobenius-orthogonal coordinates. The symmetric coordinate $\log K$ encodes the metric geometry of the head, the circulation ζ measures its deviation from pure diffusion, and the offset β functions like an attention sink by adding a query-independent bias to each key. We also showed that a softmax head implements an advection-diffusion step that is always contractive on features, and that different normal coordinates control properties of smoothing and squashing on the complete graph. Empirically, it was seen that softmax heads in pretrained encoders become more Laplacian-like with depth, and that token-centering removes β exactly, consistent with theoretical predictions. Note that the normal form treats one head at a time: except for Corollary 4.4, the dynamical results do not account for value projections, normalization layers, or multihead mechanisms. Some results also hold only for the complete graph, which limits their applicability in all scenarios of interest. Future work will address these limitations.

AI USE STATEMENT

This work made substantial use of generative AI tools, namely Claude Opus 5 and ChatGPT Codex. While the core results and ideas are due to the authors, AI has assisted in mathematical proof verification and typesetting, code and figure generation, literature review and identification, preliminary ideation (i.e., brainstorming) of mathematical and experimental content, manuscript editing and internal consistency checking. Generative AI tools were **not** used for conceptualization, methodology, story-crafting, drawing conclusions from theory/experiments, main draft writing, contextualization of results within the broader literature, and other "higher-level" objectives necessary for good research. The authors take full responsibility for the final content of this work, including that produced with the aid of generative AI.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Petar Veličković, Razvan Pascanu, and Michael M. Bronstein. Why do LLMs attend to the first token? In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=tu4dFUsW5z.

Ben Chamberlain, James Rowbottom, Maria I. Gorinova, Michael Bronstein, Stefan Webb, and Emanuele Rossi. GRAND: Graph neural diffusion. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1407–1418, 2021a.

Benjamin Paul Chamberlain, James Rowbottom, Davide Eynard, Francesco Di Giovanni, Xiaowen Dong, and Michael M. Bronstein. Beltrami flow and neural diffusion on graphs. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1594–1609, 2021b.

Jeongwhan Choi, Seoyoung Hong, Noseong Park, and Sung-Bae Cho. GREAD: Graph neural reaction-diffusion networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5722–5747, 2023.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. doi: 10.1016/j.acha.2006.04.006.

Gabriele Corso, Hannes Stark, Stefanie Jegelka, Tommi Jaakkola, and Regina Barzilay. Graph neural networks. *Nature Reviews Methods Primers*, 4(1):17, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186, 2019.

Francesco Di Giovanni, Lorenzo Giusti, Federico Barbero, Giulia Luise, Pietro Liò, and Michael M. Bronstein. On over-squashing in message passing neural networks: The impact of width, depth, and topology. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7865–7885, 2023.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2793–2803, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

James Eells and J. H. Sampson. Harmonic mappings of Riemannian manifolds. *American Journal of Mathematics*, 86(1):109–160, 1964. doi: 10.2307/2373037.

Moshe Eliasof, Eldad Haber, and Eran Treister. Feature transportation improves graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11874–11882, 2024. doi: 10.1609/aaai.v38i11.29073.

Yotam Erel, Olaf Dünkel, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Amit H. Bermano. Attention (as discrete-time Markov) chains. In *Advances in Neural Information Processing Systems*, 2025.

Theodore Frankel. *The Geometry of Physics: An Introduction*. Cambridge University Press, Cambridge, third edition, 2012. ISBN 9781107602601. doi: 10.1017/CBO9781139061377.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on Transformers. *Bulletin of the American Mathematical Society*, 62(3):427–479, 2025. doi: 10.1090/bull/1863.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016.

Leo J. Grady and Jonathan R. Polimeni. *Discrete Calculus: Applied Analysis on Graphs for Computational Science*. Springer, London, 2010.

Alexander Grigor'yan. *Introduction to Analysis on Graphs*, volume 71 of *University Lecture Series*. American Mathematical Society, Providence, RI, 2018.

Anthony Gruber, Kookjin Lee, and Nathaniel Trask. Reversible and irreversible bracket-based dynamics for deep graph neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 38454–38484, 2023.

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 (ICLR)*, 2025.

John David Jackson. *Classical Electrodynamics*. Wiley, New York, third edition, 1999. ISBN 047130932X.

Matthias Keller, Daniel Lenz, and Radosław K. Wojciechowski. *Graphs and Discrete Dirichlet Spaces*, volume 358 of *Grundlehren der mathematischen Wissenschaften*. Springer, Cham, 2021. doi: 10.1007/978-3-030-81459-5.

Frank P. Kelly. *Reversibility and Stochastic Networks*. Wiley, Chichester, 1979.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

A. Kolmogoroff. Zur Theorie der Markoffschen Ketten. *Mathematische Annalen*, 112(1):155–160, 1936. doi: 10.1007/BF01565412.

Wonsuk Lee. On the invariants of softmax attention. *arXiv preprint arXiv:2605.02907*, 2026.

David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, Providence, RI, second edition, 2017. With contributions by Elizabeth L. Wilmer.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3538–3545, 2018. doi: 10.1609/aaai.v32i1.11604.

Lek-Heng Lim. Hodge Laplacians on graphs. *SIAM Review*, 62(3):685–715, 2020. doi: 10.1137/18M1223101.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Mahdi Namazifar, Devamanyu Hazarika, and Dilek Hakkani-Tür. Role of bias terms in dot-product attention. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations (ICLR)*, 2020.

Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3515–3530, 2022.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling (COLM)*, 2024.

Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M. Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. In *International Conference on Learning Representations (ICLR)*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007. doi: 10.1007/s11222-007-9033-z.

Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the Fourier domain analysis: From theory to practice. In *International Conference on Learning Representations (ICLR)*, 2022.

Qitian Wu, Chenxiao Yang, Wentao Zhao, Yixuan He, David Wipf, and Junchi Yan. DIFFormer: Scalable (graph) transformers induced by energy constrained diffusion. In *International Conference on Learning Representations (ICLR)*, 2023a.

Xinyi Wu, Amir Ajorlou, Zihui Wu, and Ali Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

Kai Zhao, Qiyu Kang, Yang Song, Rui She, Sijie Wang, and Wee Peng Tay. Graph neural convection-diffusion with heterophily. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*, pp. 4656–4664, 2023. doi: 10.24963/ijcai.2023/518.

A Harmonic Maps and the Dirichlet Energy

This Appendix records some useful facts about the continuous Dirichlet energy and its discrete counterpart used in the body.

Consider Riemannian manifolds (M, g) and (N, h) in the continuous setting. The smoothness of a mapping $f : M \to N$ is often quantified by its *Dirichlet energy* (Eells & Sampson, 1964)

$$\mathcal{D}(f) = \frac{1}{2}\int_M \operatorname{tr}_g(f^*h)\,\mu_g = \frac{1}{2}\int_M |\mathrm{d}f|^2_{g^{-1}\otimes f^*h}\,\mu_g, \tag{3}$$

where μ_g is the volume form on M, g^{-1} is the induced metric on the cotangent bundle T^*M (i.e., co-metric on M), and $(f^*h)(u,v) = h(\mathrm{d}f(u), \mathrm{d}f(v))$ denotes the pullback metric on M along f. Here, the differential $\mathrm{d}f \in T^*M \otimes f^*TN$ is a (1,1)-tensor field, i.e., a smoothly varying vector-valued one-form $\mathrm{d}f_x \in T_x^*M \otimes T_{f(x)}N$ at each point $x \in M$. Critical points of $\mathcal{D}$, known as *harmonic maps*, represent the "locally least tense" ways of stretching M onto N. For example, given an unstretched rubber band M and a smooth stone N, a harmonic map is an application of the rubber band around the stone in such a way that no local perturbation to the band can decrease its overall elastic energy. This is characterized by the vanishing of the *tension field* $\tau(f) = \mathrm{tr}_g \nabla \mathrm{d}f$ where $\nabla : \Gamma(f^*TN) \to \Gamma(T^*M \otimes f^*TN)$ is the pullback of the Levi-Civita covariant derivative on N.

Proposition A.1. *Critical points of the Dirichlet energy* (3) *satisfy* $\tau(f) = \mathrm{tr}_g \nabla \mathrm{d}f = 0$.

Proof. Consider any compactly supported variation $\mathring{f} : M \to N$ of the mapping $f : M \to N$. It is useful to consider $\mathring{f} = \partial_\varepsilon f$ to be the derivative of f, as a one-parameter family of mappings $f(\cdot, \varepsilon)$, in the direction of an independent coordinate ε. It follows that the derivative of $\mathcal{D}$ in the direction $\mathring{f}$ is given by

$$\begin{aligned}\mathring{\mathcal{D}}(f) &= \frac{d}{d\varepsilon}\Big|_{\varepsilon=0} \frac{1}{2} \int_M \mathrm{tr}_g(f^*h)\, \mu_g = \int_M \mathrm{tr}_g \langle \mathrm{d}f, \nabla_{\partial_\varepsilon} \mathrm{d}f \rangle_h \, \mu_g \\ &= \int_M \mathrm{tr}_g \langle \mathrm{d}f, \nabla \mathring{f} \rangle_h \, \mu_g = \int_M \big\langle \mathrm{tr}_g \nabla \mathrm{d}f, \mathring{f} \big\rangle_h \, \mu_g,\end{aligned}$$

where the penultimate equality follows from torsion-freeness of ∇ and the final equality applies integration by parts. More precisely, torsion-freeness implies that the Hessian $\nabla \mathrm{d}f \in \Gamma(T^*M \otimes T^*M \otimes f^*TN)$ is symmetric in its first two indices, yielding that

$$\nabla \mathrm{d}f(u, \partial_\varepsilon) - \nabla \mathrm{d}f(\partial_\varepsilon, u) = \nabla_u \mathring{f} - \nabla_{\partial_\varepsilon} \mathrm{d}f(u) = 0.$$

Therefore, the variation $\mathring{\mathcal{D}}$ of the Dirichlet energy vanishes for all admissible $\mathring{f}$ only when the tension field $\tau(f) = \mathrm{tr}_g \nabla \mathrm{d}f = 0$ vanishes identically. □

The prototypical example of Proposition A.1 is given by the usual Dirichlet energy of real-valued mappings $u : \mathbb{R}^2 \to \mathbb{R}$ in the Euclidean geometry, in which case the tension field is simply the (negative) Laplace operator $\tau(u) = -\Delta u$. Given a graph $\mathcal{G}$ with geometry (M, m), a feature matrix $F \in \mathbb{R}^{V \times n}$, and a metric tensor $N = N(F) \in \mathbb{R}^{n \times n}$ on $\mathbb{R}^n$, there is a discrete Dirichlet energy defined analogously to the continuous one

$$\mathcal{D}_{M,N}(F) \;=\; \tfrac{1}{2}\,\mathrm{tr}_M(F^*N) \;=\; \tfrac{1}{2} \sum_{i<j} M_{ij}\, |\vec{f_i} - \vec{f_j}|_N^2 \;=:\; \tfrac{1}{2} |\mathrm{d}F|_{M \otimes N}^2. \tag{4}$$

Applied to mappings into $\mathbb{R}$ with $N = 1$, i.e., feature vectors $f \in \mathbb{R}^V$, this becomes the usual object

$$\mathcal{D}_M(f) \;=\; \tfrac{1}{2} \sum_{i<j} M_{ij} (f_i - f_j)^2 \;=\; \tfrac{1}{2} |\mathrm{d}f|_M^2.$$

Note that the Dirichlet energy depends only on the metric data (M, N) and not on the node volume m.

Harmonicity is characterized in the continuous setting by the vanishing of the *tension field* $\tau(f) = \mathrm{tr}_g \nabla \mathrm{d}f$, where $\nabla : \Gamma(f^*T\mathcal{N}) \to \Gamma(T^*\mathcal{M} \otimes f^*T\mathcal{N})$ is the pullback of the Levi-Civita connection on $\mathcal{N}$. When the target metric N is constant (i.e., the co-domain is flat), the discrete tension field takes the familiar form of the metric graph Laplacian.

Proposition A.2. *For constant N, the discrete Dirichlet energy can be computed from the (not necessarily compatible) metric Laplacian,*

$$\mathcal{D}_{M,N}(F) = \tfrac{1}{2} \langle F, \Delta_{M,m} F \rangle_{m \otimes N},$$

Since $\Delta_{M,m}$ is m-self-adjoint, the $(m \otimes N)$-gradient of $\mathcal{D}_{M,N}$ is $\Delta_{M,m}F$, and the discrete tension field is

$$\tau_{M,m}(F) = -\Delta_{M,m} F, \qquad \tau_{M,m}(F)_i = \frac{1}{m_i} \sum_j M_{ij} (\vec{f_j} - \vec{f_i}).$$

It follows that decreasing the Dirichlet energy $\mathcal{D}_{M,N}(F)$ is equivalent to smoothing the map F. Moreover, all critical points are constant due to the constancy of N.

Example. The so-called normalized Dirichlet energy of $f \in \mathbb{R}^V$, given by

$$\mathcal{D}_{\mathrm{norm}}(f) = \frac{1}{2}\sum_{i<j} A_{ij}\left(\frac{f_j}{\sqrt{d_j}} - \frac{f_i}{\sqrt{d_i}}\right)^2 = \frac{1}{2}\langle f, (I - \tilde{d}^{-1/2} A \tilde{d}^{-1/2}) f\rangle,$$

satisfies $D_{\mathrm{norm}}(f) = \mathcal{D}_A(\tilde{d}^{1/2} f)$ in terms of the discrete Dirichlet energy. That is, it differs only through a choice of gauge.

In the case of attention, the edge conductances $M = M(F)$ depend on the map, so the corresponding tension field picks up another term.

Proposition A.3. *Let $M = M(F)$ be symmetric, $m = m(F)$ be positive, and N be constant. Let $M'(F) : \mathbb{R}^{V\times n} \to \mathbb{R}^{V\times V}_{\mathrm{sym}}$ denote the Fréchet derivative with adjoint $M'(F)^\dagger : \mathbb{R}^{V\times V}_{\mathrm{sym}} \to \mathbb{R}^{V\times n}$ compatible with the $(m \otimes N)$-pairing on nodes and the half-Frobenius pairing on edges. Then, the tension field is given by*

$$\tau_{M,m}(F) = -\Delta_{M,m} F - \tfrac{1}{2} M'(F)^\dagger (F^* N).$$

In the case of the scaled dot product attention, this becomes the following.

Corollary A.4. *Suppose $M(F) = \exp(FSF^\intercal)$ for symmetric S, N is constant, and $m = M1$ is specified by compatibility. Then, the tension field is given by*

$$\tau_{M,m}(F) = -\Delta_{M,m} F - \tfrac{1}{2}(A \odot F^* N) F S N^{-1}, \qquad A = \tilde{m}^{-1} M.$$

Observe that an update based on $\mathcal{D}_{M,N}$ is different from what is applied by a typical attention layer $F \mapsto AFW$ due to the variation of the edge conductance.

B Omitted Statements and Proofs

Here we collect proofs of theoretical results contained in the body, along with auxiliary statements necessary for the main results.

B.1 Section 2

The following explicit form of the metric graph Laplacian is known (see Gruber et al. (2023); Grigor'yan (2018)) and useful for calculations.

Proposition B.1. *The metric graph Laplacian $\Delta_{M,m} f = \mathrm{d}^*_{M,m} \mathrm{d} f$ of a feature vector $f \in \mathbb{R}^V$ is the m-self-adjoint operator $\Delta_{M,m} f = \tilde{m}^{-1} L_M f = m^{-1} \odot \mathrm{d}^\intercal (M \odot \mathrm{d} f)$ where $\mathrm{d}^*_{M,m} : \Omega^1(\mathcal{G}) \to \mathbb{R}^V$ is the metric graph divergence given by $\mathrm{d}^*_{M,m} U = m^{-1} \odot \mathrm{d}^\intercal (M \odot U)$.*

Proof of Proposition B.1. For any $f \in \mathbb{R}^V$ and $U \in \Omega^1(\mathcal{G})$, observe that

$$\langle \mathrm{d} f, U\rangle_M = \langle f, \mathrm{d}^\intercal (M \odot U)\rangle = \langle f, m^{-1} \odot \mathrm{d}^\intercal (M \odot U)\rangle_m = \langle f, \mathrm{d}^*_{M,m} U\rangle_m.$$

Therefore, $\mathrm{d}^*_{M,m} U = m^{-1} \odot \mathrm{d}^\intercal (M \odot U)$ and $\Delta_{M,m} f = \mathrm{d}^*_{M,m} \mathrm{d} f = m^{-1} \odot \mathrm{d}^\intercal (M \odot \mathrm{d} f)$ as claimed. Moreover, observe that

$$[\Delta_{M,m} f]_i = \frac{1}{m_i} \sum_{j\sim i} M_{ij}(f_i - f_j) = m_i^{-1} \sum_{j\sim i} [\mathrm{Diag}(M1) - M]_{ij} f_j = [\tilde{m}^{-1} L_M f]_i,$$

establishing the result. □

Proof of Corollary 2.1. Using the formula $\Delta_{M,m} = \tilde{m}^{-1} L_M$ from Proposition B.1, it follows that $\Delta_{M,m} = \tilde{m}^{-1} L_M = \tilde{m}^{-1}(\mathrm{diag}(M1) - M) = I - \tilde{m}^{-1} M$ where the last equality holds if and only if $m = M1$. Note that entrywise nonnegativity follows from positivity of M, m and row-stochasticity follows since $\tilde{m}^{-1} M1 = \tilde{m}^{-1} m = 1$. Finally, for any $f, g \in \mathbb{R}^V$ it follows that

$$\langle Af, g\rangle_m = \langle Mf, g\rangle = \langle f, Mg\rangle = \langle f, \tilde{m}^{-1} Mg\rangle_M = \langle f, Ag\rangle_m,$$

proving that A is m-self-adjoint. □

Proposition B.2. *Every family of symmetric score matrices $\{C + g1^\intercal : g \in \mathbb{R}^V\}$ corresponds to a unique softmax head $A = I - \Delta_{M,m}$ in the compatible geometry where $M = \exp C$ up to a positive scalar.*

Proof of Proposition B.2. Observe that $\mathrm{softmax}_j(C + g1^\intercal) = \mathrm{norm}_j(e^C \odot e^{g1^\intercal}) = \mathrm{norm}_j(e^C \odot e^g 1^\intercal) = \mathrm{norm}_j e^C = \mathrm{softmax}_j C$, so that the gauge term $g1^\intercal$ is removed. Since, $A = \mathrm{softmax}_j C = \tilde{m}^{-1} M$ with $M = \exp C$ and $m = M1$, the conclusion follows from Corollary 2.1. □

Proof of Theorem 2.3. Suppose that $L = \Delta_{M,m} = \tilde{m}^{-1} L_M$ (by Proposition B.1). Then, $L_{ij} = -M_{ij}/m_i$ for $i \neq j$, and therefore

$$\theta(L) = \frac{1}{2} \log \frac{L_{ij}}{L_{ji}} = \frac{1}{2} \log \frac{M_{ij}/m_i}{M_{ji}/m_j} = \frac{1}{2} \log \frac{m_j}{m_i} = \mathrm{d}\big(\tfrac{1}{2} \log m\big)_{ij},$$

establishing necessity. Conversely, suppose $\theta(L) = \mathrm{d}h$ for some $h \in \mathbb{R}^V$ and let $m = e^{2h}$. Then, $\theta(L)_{ij} = h_j - h_i = \frac{1}{2}\log(m_j/m_i)$, so that $L_{ij}/L_{ji} = m_j/m_i$. It follows that the matrix M with $M_{ij} = -m_i L_{ij}$ for $i \neq j$ and $M_{ii} = 0$ is symmetric and entrywise nonnegative. Since $L1 = 0$, it also satisfies $\mathrm{diag}(M1) - M = \tilde{m}L$, so $L = \tilde{m}^{-1} L_M = \Delta_{M,m}$ by Proposition B.1. For the final statement, note that $L = I - A$ has negative off-diagonal entries $-A_{ij}$ and $\theta(I - A) = \theta(A)$. Redefining $M_{ii} = m_i A_{ii}$ leaves L_M unchanged and gives $M = \tilde{m}A$ with $M1 = m$. Hence (M, m) is compatible and $A = I - \Delta_{M,m}$ by Corollary 2.1. □

Proof of Corollary 2.5. This follows immediately from Theorem 2.3 and the fact that every closed cochain $\alpha \in \Omega^1(\mathcal{G})$ on the complete graph $\mathcal{G}$ is also exact, i.e., $\alpha = \mathrm{d}f$ for some $f \in \mathbb{R}^V$ if and only if $\mathrm{d}\alpha = 0$. □

Proposition B.3. *Let $s \in \mathbb{R}^V_+$ be a positive vector and $f \mapsto s \odot f$ be a change of nodal variables. Then, the map $(\Delta_{M,m}, m) \mapsto \big(\tilde{s}\,\Delta_{M,m}\,\tilde{s}^{-1}, m \oslash s^2\big)$ is spectrum-preserving, and the operator $\tilde{s}\Delta_{M,m}\tilde{s}^{-1}$ is $(m \oslash s^2)$-self-adjoint with balance cochain $\mathrm{d}\log(\sqrt{m} \oslash s)$ when (M, m) is compatible. In particular, the transformed operator is only a metric laplacian when s is constant.*

Proof. Isospectrality is immediate from similarity of $\Delta_{M,m}$ and $\tilde{s}\Delta_{M,m}\tilde{s}^{-1}$. The m-self-adjointness follows from

$$\langle f, \tilde{s}\Delta_{M,m}\tilde{s}^{-1} g\rangle_{m \oslash s^2} = \langle f \oslash s, \Delta_{M,m}(g \oslash s)\rangle_m = \langle \Delta_{M,m}(f \oslash s), g \oslash s\rangle_m = \langle \tilde{s}\Delta_{M,m}\tilde{s}^{-1} f, g\rangle_{m \oslash s^2}.$$

Now, observe that compatibility of (M, m) implies

$$\theta(\tilde{s}\Delta_{M,m}\tilde{s}^{-1})_{ij} = \frac{1}{2} \log \frac{s_i^2 m_j}{s_j^2 m_i} = \frac{1}{2}\mathrm{d}\log m - \mathrm{d}\log s = \mathrm{d}\log(\sqrt{m} \oslash s),$$

so that the balance cochain is as claimed. Finally, every metric Laplacian annihilates constants since $L_M 1 = 0$, which places the restriction $\tilde{s}^{-1}1 = c^{-1}1$ for some positive constant c^{-1}, i.e., $s = c1$. □

B.2 Section 3

Proof of Proposition 3.1. The first claim is immediate from the linearity of sym and skew along with $\mathrm{skew}(g1^\intercal)_{ij} = \frac{1}{2}(g_i - g_j) = -\frac{1}{2}\mathrm{d}g$. To check gauge invariance, we compute

$$\mathrm{av}(g)_{ii} + \mathrm{av}(g)_{jj} - \mathrm{av}(g)_{ij} - \mathrm{av}(g)_{ji} = 0, \qquad \mathrm{dskew}\, g = -\tfrac{1}{2}\mathrm{d}^2 g = 0.$$

Now, observe that $\log A = a - \log(h1^\intercal)$ where $h = e^a 1$. Therefore, $\mathrm{skew}\log A = \mathrm{skew}\, a + \frac{1}{2}\mathrm{d}\log h$, and $\mathrm{d}\theta(A) = \mathrm{d}\,\mathrm{skew}\log A = \mathrm{d}\,\mathrm{skew}\, a = \xi$. The cross ratio formulas now follow from direct substitution. □

Proof of Corollary 3.2. Observe that $\mathrm{sym}\, a = FSF^\intercal$ and $\mathrm{skew}\, a = FBF^\intercal$. Therefore,

$$G_{ij} = \vec{f_i}^\intercal S \vec{f_i} + \vec{f_j}^\intercal S \vec{f_j} - 2\vec{f_i}^\intercal S \vec{f_j} = (\vec{f_i} - \vec{f_j})^\intercal S (\vec{f_i} - \vec{f_j}) = F^* S_{ij},$$
$$\xi_{ijk} = \vec{f_i}^\intercal B \vec{f_j} + \vec{f_j}^\intercal B \vec{f_k} + \vec{f_k}^\intercal B \vec{f_i} = (\vec{f_i} - \vec{f_k})^\intercal B (\vec{f_j} - \vec{f_k}) = F^* B_{ijk},$$

where the penultimate equality used that $\vec{f}_k^{\intercal} B \vec{f}_k = 0$ by antisymmetry. To prove that F^*B encodes area, parameterize $T = T_{ijk}$ as $\vec{f}_k + t_1(\vec{f}_i - \vec{f}_k) + t_2(\vec{f}_j - \vec{f}_k)$ with $t_1, t_2 \geq 0$ and $0 \leq t_1 + t_2 \leq 1$. Then, the 1-form $\alpha_{\vec{u}}(\vec{v}) = \frac{1}{2}\vec{u}^{\intercal} B \vec{v} = \frac{1}{2}\omega(\vec{u}, \vec{v})$ satisfies $d\alpha = \omega$ (since $d\big(\frac{1}{2}\sum_{a,b} u_a B_{ab}\, du^b\big) = \sum_{a<b} B_{ab}\, du^a \wedge du^b$), and so Stokes' theorem guarantees

$$\int_T \omega = \int_{\partial T} \alpha = \tfrac{1}{2}\big(\omega(\vec{f}_k, \vec{f}_i) + \omega(\vec{f}_i, \vec{f}_j) + \omega(\vec{f}_j, \vec{f}_k)\big) = \tfrac{1}{2}\xi_{ijk} = \tfrac{1}{2}F^* B_{ijk}.$$

Here ∂T is traversed as $\vec{f}_k \to \vec{f}_i \to \vec{f}_j \to \vec{f}_k$, and each segment contributes $\int_{[\vec{p},\vec{q}]} \alpha = \frac{1}{2}\int_0^1 (\vec{p} + s(\vec{q} - \vec{p}))^{\intercal} B(\vec{q} - \vec{p})\, ds = \frac{1}{2}\vec{p}^{\intercal} B \vec{q}$ by antisymmetry of B. □

Proof of Corollary 3.3. The first statement is direct from $d\theta(A) = F^*B$ and Corollary 2.5. Since $\vec{f}_j - \vec{f}_i = (\vec{f}_j - \vec{f}_k) - (\vec{f}_i - \vec{f}_k)$, the vectors $\vec{f}_i - \vec{f}_k$ span $\operatorname{span} \mathrm{d}F$ for any fixed k, so $F^*B = 0$ if and only if $\vec{u}^{\intercal} B \vec{v} = 0$ on $\operatorname{span} \mathrm{d}F$. The second statement follows since $U = \operatorname{span} dF$ is then isotropic for B, i.e., $B(U) \subseteq U^{\perp}$. By rank-nullity,

$$\dim U - (n - 2r) \leq \dim U - \dim(U \cap \ker B) = \dim B(U) \leq n - \dim U,$$

so that $\dim U \leq n - r$. □

Proof of Theorem 3.4. Consider the space $\mathcal{B} = \{u1^{\intercal} + 1v^{\intercal} : u, v \in \mathbb{R}^V\}$, and let P denote Frobenius-orthogonal projection onto $\mathcal{B}^{\perp}$. Then, $\mathcal{B}$ is invariant under transposition, hence the space $\mathbb{R}^{V\times V}/\mathcal{B} \cong (\operatorname{sym} \cap \mathcal{B}^{\perp}) + (\operatorname{skew} \cap \mathcal{B}^{\perp})$ has a symmetric/skew-symmetric decomposition. Now, if $\mathcal{G} = \{g1^{\intercal} : g \in \mathbb{R}^V\}$ denotes the gauge group of the softmax, then $\mathcal{G} \subset \mathcal{B}$ and choosing the Coulomb gauge is equivalent to picking the representative $a \in [a]$ for $[a] \in \mathbb{R}^{V\times V}/\mathcal{B}$ that satisfies $a \perp \mathcal{G}$. Letting $\mathcal{G}^{\perp_{\mathcal{B}}}$ denote the complement of $\mathcal{G}$ in $\mathcal{B}$, it follows that $\mathbb{R}^{V\times V}/\mathcal{G} \cong (\operatorname{sym} \cap \mathcal{B}^{\perp}) + (\operatorname{skew} \cap \mathcal{B}^{\perp}) + \mathcal{G}^{\perp_{\mathcal{B}}}$.

Since $\langle X, u1^{\intercal} + 1v^{\intercal}\rangle = \langle u, X1\rangle + \langle v, X^{\intercal}1\rangle$, the space $\mathcal{B}^{\perp}$ consists of the X with $X1 = X^{\intercal}1 = 0$. Hence $\operatorname{sym} \cap \mathcal{B}^{\perp}$ and $\operatorname{skew} \cap \mathcal{B}^{\perp}$ are exactly the spaces of $\log K$ and ζ in the statement, where $\zeta 1 = -\mathrm{d}^{\intercal}\zeta$. For the last summand, $\mathcal{G}^{\perp_{\mathcal{B}}} \cong \mathcal{B}/\mathcal{G}$, and every $u1^{\intercal} + 1v^{\intercal}$ is gauge-equivalent to $1v^{\intercal}$. Moreover, $1v^{\intercal} = g1^{\intercal}$ forces $g_i = v_i$ on the self-loops and $v_i = v_j$ on the edges, so $1v^{\intercal} \in \mathcal{G}$ if and only if v is constant. Therefore $\mathcal{B}/\mathcal{G} \cong \{1\beta^{\intercal} : 1^{\intercal}\beta = 0\}$ via $[u1^{\intercal} + 1v^{\intercal}] \mapsto \beta = v - \frac{1}{V}(1^{\intercal}v)1$, and every $A \in \mathcal{A} \cong \mathbb{R}^{V\times V}/\mathcal{G}$ has exactly one normal form. The coordinates are Frobenius-orthogonal because $\operatorname{im}\operatorname{sym} \perp \operatorname{im}\operatorname{skew}$ and $\langle X, 1\beta^{\intercal}\rangle = \langle X^{\intercal}1, \beta\rangle = 0$ for $X \in \mathcal{B}^{\perp}$. The map $(\log K, \zeta, \beta) \mapsto \operatorname{softmax}_j(\log K + \zeta + 1\beta^{\intercal})$ is smooth, and its inverse $A \mapsto (\operatorname{sym} P \log A, \operatorname{skew} P \log A, \beta)$ is linear in $\log A$, so it is a diffeomorphism onto $\mathcal{A}$. For the dimensions, $X \mapsto X1$ maps the $(E + V)$-dimensional space of symmetric matrices onto $\mathbb{R}^V$ (diagonal matrices suffice), and $\mathrm{d}^{\intercal}$ maps $\Omega^1(G)$ onto $(\ker \mathrm{d})^{\perp} = 1^{\perp}$ because G is connected. Hence $\dim \log K = E$, $\dim \zeta = E - V + 1$, and $\dim \beta = V - 1$. □

Proof of Corollary 3.5. Observe that $K \odot 1b^{\intercal}$ and $K \odot bb^{\intercal}$ differ by a row scaling in the kernel of norm_j, so that $A = \operatorname{softmax}_j(\log K + \zeta + 1\beta^{\intercal}) = \operatorname{norm}_j(M \odot e^{\zeta})$ as claimed. The second statement is immediate from Corollary 2.1 when $\zeta = 0$, with $m = M1$. Conversely, if A is the complement of a metric Laplacian, then $\theta(A)$ is exact by Theorem 2.3. Since $\log A = \log K + \zeta + 1\beta^{\intercal} + g1^{\intercal}$ for some $g \in \mathbb{R}^V$, it follows that $\theta(A) = \operatorname{skew} \log A = \zeta + \frac{1}{2}\mathrm{d}(\beta - g)$, so $\zeta = \mathrm{d}h$ is also exact. Because $\mathrm{d}^{\intercal}\zeta = 0$, this gives $|\zeta|^2 = \langle \mathrm{d}h, \zeta\rangle = \langle h, \mathrm{d}^{\intercal}\zeta\rangle = 0$. □

Proof of Corollary 3.6. Since $\mathcal{G}$ is complete, the Coulomb representative of a is aP with $P = I - \frac{1}{V}11^{\intercal}$, and column-centering gives

$$\log K + \zeta = PaP = HWH^{\intercal}, \qquad 1\beta^{\intercal} = aP - PaP = \tfrac{1}{V}11^{\intercal}aP,$$

so that $\log K = HSH^{\intercal}$, $\zeta = HBH^{\intercal}$, and $\beta = \frac{1}{V}PFW^{\intercal}F^{\intercal}1 = HW^{\intercal}\bar{f}$. Since $P1 = 0$ and $F^*S = s1^{\intercal} + 1s^{\intercal} - 2FSF^{\intercal}$ with $s_i = |\vec{f}_i|_S^2$, it follows that $-\frac{1}{2}P(F^*S)P = HSH^{\intercal}$. Similarly, $\mathrm{d}^{\intercal}$ is the transposed incidence matrix, so $[\mathrm{d}^{\intercal}F^*B]_{ij} = \sum_k F^*B_{ijk}$, and

$$\tfrac{1}{V}\textstyle\sum_k (\vec{f}_i - \vec{f}_k)^{\intercal} B(\vec{f}_j - \vec{f}_k) = \vec{f}_i^{\intercal} B \vec{f}_j - \vec{f}_i^{\intercal} B \bar{f} - \bar{f}^{\intercal} B \vec{f}_j = (\vec{f}_i - \bar{f})^{\intercal} B(\vec{f}_j - \bar{f}) = [HBH^{\intercal}]_{ij}.$$

Finally, $H = PF$ is invariant under translations of F. Its rows span $\operatorname{span}\{\vec{f}_i - \bar{f}\} = \operatorname{span} \mathrm{d}F$, so $\zeta = HBH^{\intercal} = 0$ if and only if $F^*B = 0$ by Corollary 3.3. The claim about β is immediate. □

B.3 SECTION 4

Consider the edge averaging operator $\mathrm{av} : \mathbb{R}^V \to \mathcal{S}^1(\mathcal{G})$ that acts via $[\mathrm{av}\, f]_{ij} = \frac{1}{2}(f_i + f_j)$. The following result records some properties necessary for the remaining proofs.

Proposition B.4. *The edge averaging operator obeys a Leibniz rule* $\mathrm{d}(fg) = \mathrm{d}f \odot \mathrm{av}\, g + \mathrm{av}\, f \odot \mathrm{d}g$ *and its* (M, m)*-adjoint applied to* $S \in \mathcal{S}^1(\mathcal{G})$ *is given by*

$$\mathrm{av}^*_{M,m}\, S = m^{-1} \odot \mathrm{av}^\intercal(M \odot S), \qquad [\mathrm{av}^\intercal\, S]_i = \tfrac{1}{2}\textstyle\sum_j S_{ij}.$$

Moreover, if $u \in \Omega^1(\mathcal{G})$ *is a 1-cochain, the adjoint obeys the product rule*

$$\mathrm{av}^*_{M,m}\big(u \odot \mathrm{d}F\big) = \big(\mathrm{d}^*_{M,m} u\big) \odot F - \mathrm{d}^*_{M,m}\big(u \odot \mathrm{av}\, F\big).$$

Proof of Proposition B.4. The Leibniz rule is direct, since $\frac{1}{2}(f_j - f_i)(g_i + g_j) + \frac{1}{2}(f_i + f_j)(g_j - g_i) = f_j g_j - f_i g_i$. For the adjoint, symmetry of $M \odot S$ gives

$$\langle \mathrm{av} f, S\rangle_M = \tfrac{1}{2}\sum_{i<j} M_{ij} S_{ij}(f_i + f_j) = \tfrac{1}{2}\sum_i f_i \sum_j M_{ij} S_{ij} = \big\langle f, m^{-1} \odot \mathrm{av}^\intercal(M \odot S)\big\rangle_m .$$

For the product rule, the ith rows of $(\mathrm{d}^*_{M,m} u) \odot F$ and $-\mathrm{d}^*_{M,m}(u \odot \mathrm{av} F)$ are $-\frac{1}{m_i}\sum_j M_{ij} u_{ij} \vec{f}_i$ and $\frac{1}{m_i}\sum_j M_{ij} u_{ij} \frac{1}{2}(\vec{f}_i + \vec{f}_j)$, which sum to $\frac{1}{m_i} \cdot \frac{1}{2}\sum_j M_{ij} u_{ij}(\vec{f}_j - \vec{f}_i) = [\mathrm{av}^*_{M,m}(u \odot \mathrm{d}F)]_i$. □

Proof of Theorem 4.1. By Corollary 3.5, $A = \tilde{m}^{-1}(M \odot e^\zeta)$, and $M \odot e^\zeta = M \odot (\cosh\zeta + \sinh\zeta) = \bar{M} + \frac{1}{2}\bar{M} \odot u$. Since $A1 = 1$, it follows that

$$[(A - I)F]_i = \sum_j A_{ij}(\vec{f}_j - \vec{f}_i) = \frac{1}{m_i}\sum_j \bar{M}_{ij}(\vec{f}_j - \vec{f}_i) + \frac{1}{m_i} \cdot \frac{1}{2}\sum_j \bar{M}_{ij} u_{ij}(\vec{f}_j - \vec{f}_i),$$

and Propositions B.1 and B.4 identify the two terms with $-\Delta_{\bar{M},m} F$ and $\mathrm{av}^*_{\bar{M},m}(u \odot \mathrm{d}F)$. Moreover, $\bar{M} \odot u$ is skew-symmetric and $\mathrm{d}^*_{\bar{M},m} u = -m^{-1} \odot (\bar{M} \odot u)1$, so that

$$m = \big(\bar{M} + \tfrac{1}{2}\bar{M} \odot u\big)1 = \bar{M}1 - \tfrac{1}{2} m \odot \mathrm{d}^*_{\bar{M},m} u, \qquad A^\intercal m = \big(\bar{M} - \tfrac{1}{2}\bar{M} \odot u\big)1 = \bar{M}1 + \tfrac{1}{2} m \odot \mathrm{d}^*_{\bar{M},m} u,$$

which rearrange to the stated identities. □

Proof of Corollary 4.2. By Theorem 4.1, $A_{ij} = \bar{M}_{ij}(1 + \frac{1}{2}u_{ij})/m_i$, which is positive because $|u_{ij}| < 2$. For $h \in (0, 1]$, the forward Euler step $F \mapsto \big((1 - h)I + hA\big)F$ therefore replaces each token by a convex combination of tokens. For convex weights λ, μ,

$$\Big|\sum_j \lambda_j \vec{f}_j - \sum_k \mu_k \vec{f}_k\Big| \le \sum_{j,k} \lambda_j \mu_k |\vec{f}_j - \vec{f}_k| \le \delta(F),$$

so no step increases δ. The Euler iterates with step $h = t/N$ converge to $F(t)$ as $N \to \infty$, and δ is continuous, so $\delta(F(t)) \le \delta(F_0)$. The argument is unchanged when $A = A(F)$ depends on the features. □

Corollary B.5. *A scaled dot-product head* $A \simeq (\log K, \zeta, \beta)$ *on a complete graph* $\mathcal{G}$ *has velocity cochain* $u = 2\tanh(HBH^\intercal)$*. Letting* $Y = \log K + \zeta = HWH^\intercal$ *and* $b = e^\beta$*, the diffusion and advection terms are given by*

$$\Delta_{\bar{M},m} F = \tilde{m}^{-1}\mathrm{d}^\intercal\big(bb^\intercal \odot \mathrm{sym}\, e^Y \odot \mathrm{d}F\big), \quad \mathrm{av}^*_{\bar{M},m}(u \odot \mathrm{d}F) = \tilde{m}^{-1}\,\mathrm{av}^\intercal\,\big(2\, bb^\intercal \odot \mathrm{skew}\, e^Y \odot \mathrm{d}F\big),$$

and the divergence of the velocity is $\mathrm{d}^*_{\bar{M},m} u = m^{-1} \odot b \odot (e^{Y^\intercal} - e^Y) b$*. In particular,* u *is divergence-free if and only if* $e^Y b = e^{Y^\intercal} b$*.*

Proof of Corollary B.5. By Corollary 3.6, $\zeta = HBH^\intercal$ and $\log K + \zeta = HWH^\intercal$. Since $\log K$ is symmetric and ζ is skew-symmetric, $\mathrm{sym}\, e^Y = K \odot \cosh\zeta$ and $\mathrm{skew}\, e^Y = K \odot \sinh\zeta$. Hence $\bar{M} = bb^\intercal \odot \mathrm{sym}\, e^Y$ and $\bar{M} \odot u = 2\, bb^\intercal \odot \mathrm{skew}\, e^Y$, which gives the diffusion and advection terms by Theorem 4.1. Similarly,

$$\mathrm{d}^*_{\bar{M},m} u = -m^{-1} \odot (\bar{M} \odot u)1 = m^{-1} \odot \big(bb^\intercal \odot (e^{Y^\intercal} - e^Y)\big)1 = m^{-1} \odot b \odot \big(e^{Y^\intercal} b - e^Y b\big),$$

which vanishes if and only if $e^Y b = e^{Y^\intercal} b$ since $b, m > 0$. □

Proof of Theorem 4.3. Since $\mathcal{G}$ is complete, $A_{ij} = e^{Y_{ij}} b_j / \sum_k e^{Y_{ik}} b_k$, so that

$$\frac{A_{ij}}{\hat{b}_j} = \frac{e^{Y_{ij}}}{\sum_k \hat{b}_k e^{Y_{ik}}}.$$

The denominator is a $\hat{b}$-weighted average of $e^{Y_{ik}}$, so it lies between $e^{\min_k Y_{ik}}$ and $e^{\max_k Y_{ik}}$, and the bounds on A_{ij} follow. In particular, $A - e^{-\kappa(Y)} 1\hat{b}^\intercal$ is entrywise nonnegative with row sums $1 - e^{-\kappa(Y)}$. Since $1\hat{b}^\intercal F$ has identical rows, $(AF)_i - (AF)_{i'}$ is $1 - e^{-\kappa(Y)}$ times a difference of two convex combinations of tokens. As in the proof of Corollary 4.2, this gives $\delta(AF) \le (1 - e^{-\kappa(Y)})\delta(F)$. Finally, $Y = 0$ gives $A = 1\hat{b}^\intercal$ by the bounds. Conversely, $1\hat{b}^\intercal = \mathrm{softmax}_j(1\beta^\intercal)$ has normal coordinates $(0, 0, \beta)$, so $A = 1\hat{b}^\intercal$ forces $Y = 0$ by uniqueness in Theorem 3.4. □

Proof of Corollary 4.4. By Corollary 3.6, $Y = HWH^\intercal$, and $\bar{f}$ is a convex combination of tokens, so that

$$\mathrm{osc}_j\, Y_{ij} = \max_{j,k} (\vec{f}_i - \bar{f})^\intercal W (\vec{f}_j - \vec{f}_k) \le \|W\| |\vec{f}_i - \bar{f}| \delta(F) \le \|W\| \delta(F)^2.$$

Similarly, $\zeta_{ij} = \frac{1}{V} \sum_k \vec{u}^\intercal B \vec{v}$ with $\vec{u} = \vec{f}_i - \vec{f}_k$ and $\vec{v} = \vec{f}_j - \vec{f}_k$. Since $\vec{u}^\intercal B \vec{u} = 0$, only the component $\vec{v}_\perp$ of $\vec{v}$ orthogonal to $\vec{u}$ contributes, so $|\vec{u}^\intercal B \vec{v}| \le \|B\| |\vec{u}| |\vec{v}_\perp| = 2\|B\| \,\mathrm{area}(\vec{f}_i, \vec{f}_j, \vec{f}_k) \le 2\|B\| \,\mathrm{area}(F)$. For the networks, $\delta(\sigma(X)) \le L_\sigma \delta(X)$ and $\delta(XW_V) \le \|W_V\| \delta(X)$, so Theorem 4.3 and the bound on $\kappa(Y)$ give

$$\delta\big(\sigma(A(F)FW_V)\big) \le \rho\big(1 - e^{-\kappa(Y)}\big)\delta(F) \le \rho\big(1 - e^{-\|W\|\delta(F)^2}\big)\delta(F).$$

This is the first inequality, and the second follows from $|\delta(F + Z) - \delta(F)| \le \delta(Z)$. Since $1 - e^{-x} \le x$, both right-hand sides are at most $c\,\delta^3$ with $c = \rho\|W\|$. Without a residual connection, $x_m = \sqrt{c}\,\delta^{(m)}$ then satisfies $x_{m+k} \le x_m^{3^k}$, which vanishes doubly exponentially when $x_m < 1$. With a residual connection, $|\delta_{\ell+1} - \delta_\ell| \le c\,\delta_\ell^3$. □

C Additional Experiments

Here we collect some additional experiments. Full implementation details, checkpoint revisions, and additional head-resolved results are in Appendix D.

C.1 Omitted Figures

The following Figures were referenced in the body of the work.

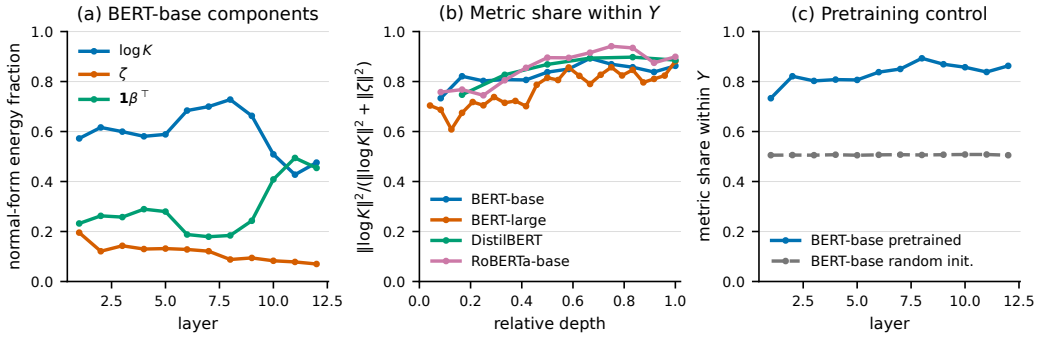

Figure 2: Normal-form energy fraction as a function of depth. (a) Orthogonal energy fractions for pretrained BERT-base. (b) Metric share r_{met} across pretrained encoders. (c) Pretrained BERT-base versus an architecture-matched random initialization. Shading denotes head-averaged 95% confidence intervals across sequences.

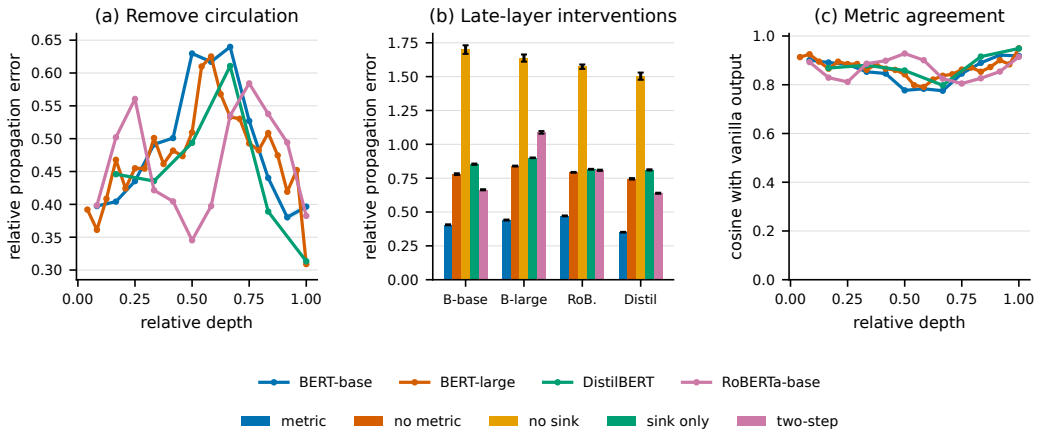

Figure 3: Results of intervening on normal form components. (a) Error after removing circulation ($\zeta = 0$) as a function of depth. (b) Late-layer output error for metric ($\zeta = 0$), no metric ($\log K = 0$), no sink ($\beta = 0$), sink only ($\log K = \zeta = 0$), and two successive applications of the metric head. (c) Cosine agreement between trained network propagation and its metric projection ($\zeta = 0$). The first legend identifies models in (a,c), and the second identifies interventions in (b). Error bars/shading are 95% confidence intervals across sequences. Late-layer ranges are defined in Table 2. Models are not retrained after intervention.

C.2 Exact gauge removal

As discussed in Namazifar et al. (2023), the gauge in the score function also identifies a small but exact learned-parameter redundancy. Writing $q_i = W_Q h_i + b_Q$ and $k_j = W_K h_j + b_K$, the key-bias contribution $q_i^\top b_K$ is constant in j and therefore eliminated by the softmax. Therefore, setting every key-projection bias to zero leaves the attention operator unchanged in exact arithmetic. This removes 9,216 parameters from BERT-base, 24,576 from BERT-large, 9,216 from RoBERTa-base, and 4,608 from DistilBERT, without post-training optimization. While this removes a relatively small amount of encoder parameters (0.0069–0.0084%), it is an exact consequence of the normal form Theorem 3.4.

To test inference rather than only head outputs, we also evaluate the public fine-tuned DistilBERT SST-2 checkpoint on all 872 validation examples. Deleting all 4,608 key biases preserves every prediction (maximum absolute logit change 1.8×10^{-5}), as expected. Interestingly, removing the entire coordinate β from every head in the first two layers changes accuracy from 91.06% to 91.40%, but removal from all six layers reduces it to 87.04% (Table 4).

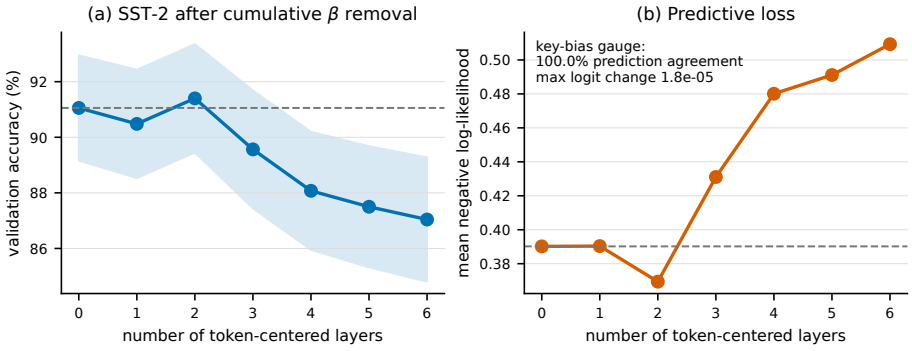

Figure 4: SST-2 inference under cumulative removal of the normal-form sink coordinate β by token-centering. The dashed line is the unmodified checkpoint. Removing the learned key biases preserves all predictions.

Table 4: SST-2 validation inference under exact gauge removal and cumulative β removal in the six-layer fine-tuned DistilBERT model. Intervals are 95% bootstraps over the 872 validation examples. Accuracy change is intervention accuracy minus baseline accuracy in percentage points (pp), and agreement is with the baseline predictions.

Intervention	Accuracy (%)	Accuracy change (pp)	Prediction agreement (%)
None (baseline)	91.06 ± 1.89	—	100.00
Remove all key biases	91.06 ± 1.89	0.00	100.00
Remove β, layers 1–2	91.40 ± 1.95	$+0.34$	98.28
Remove β, layers 1–6	87.04 ± 2.24	-4.01	89.79

C.3 Comparison of Dirichlet energy versus diameter

To test the claim in Section 4 that Dirichlet energy can increase while diameter always decreases, we consider a 64-layer GNN with residual connection, trained on the fixed Planetoid Cora split. We generate three model checkpoints that differ in their propagation operator: row-stochastic softmax attention A, the higher-order polynomial $A + A^2$, and the negative metric Laplacian $-\Delta_{M,m}$. Here M is the symmetric learned edge conductance, $m = M\mathbf{1}$, $A = \mathrm{diag}(m)^{-1}M$, and $\Delta_{M,m} = \mathrm{diag}(m)^{-1}(\mathrm{diag}(m) - M) = I - A$. Taking the initial latent features and query/key projections from each of the three trained checkpoints, we compare the update $F^{\ell+1} = A(F^\ell)F^\ell$ with the relaxed Laplacian step $F^{\ell+1} = F^\ell - \tau\Delta_{M(F^\ell),m(F^\ell)}F^\ell$, $\tau = 0.1$. We measure $\delta(F) = \max_{i,j} |\vec{f}_i - \vec{f}_j|$ and $E_\ell = D_{M(F^\ell)}(F^\ell) = \frac{1}{2}\sum_{i<j} M_{ij}(F^\ell)|\vec{f}_i^\ell - \vec{f}_j^\ell|^2$, alongside the fixed-reference energy $D_{M(F^0)}(F^\ell)$ on the same states.

The results are displayed in Figure 5. All six adaptive trajectories have non-increasing diameter. However, softmax attention increases E_ℓ at six of 64 steps for seed 0, ending at 8.85×10^3 times its initial energy while its diameter is only 0.982 of its initial value (Figure 5a,b). This small decrease in overall diameter is consistent with Corollary 4.4. The other two softmax attention trajectories and all three relaxed trajectories have decreasing Dirichlet energy, indicating that feature dependence permits Dirichlet energy growth but does not require non-monotonic decrease.

D Implementation details

Here we record various implementation details necessary for reproducing the experimental results. All experiments ran with Python 3.11.15 and PyTorch 2.11.0 on one NVIDIA H100 80GB GPU. The transformer runs additionally used Transformers 4.57.6 and Datasets 5.0.1, while the graph baselines used PyTorch Geometric 2.7.0.

Models and data. We use the public Hugging Face checkpoints listed in Table 5, without task fine-tuning. Models are put in evaluation mode, so dropout is disabled. We form candidate examples by concatenating contiguous non-heading lines of the WikiText-2 test split until each block contains at least 220 whitespace-delimited words. We sample 96 candidate blocks without replacement using seed 1729, sort the sampled source indices, tokenize with each model's native tokenizer, and pad or truncate to 128 tokens. Padding is removed before every matrix decomposition. The random control instantiates BERT-base from its configuration while retaining the same tokenizer and inputs. The sample manifest, including source-line bounds and SHA-256 hashes, is stored with the results.

Table 5: Pretrained checkpoints used in the attention experiments. Revisions are the exact Hugging Face commit hashes recorded at execution time.

Checkpoint	Layers	Heads	Revision
`google-bert/bert-base-uncased`	12	12	`86b5e0934494bd15c9632b12f734a8a67f723594`
`google-bert/bert-large-uncased`	24	16	`6da4b6a26a1877e173fca3225479512db81a5e5b`
`FacebookAI/roberta-base`	12	12	`e2da8e2f811d1448a5b465c236feacd80ffbac7b`
`distilbert/distilbert-base-uncased`	6	12	`12040accade4e8a0f71eabdb258fecc2e7e948be`

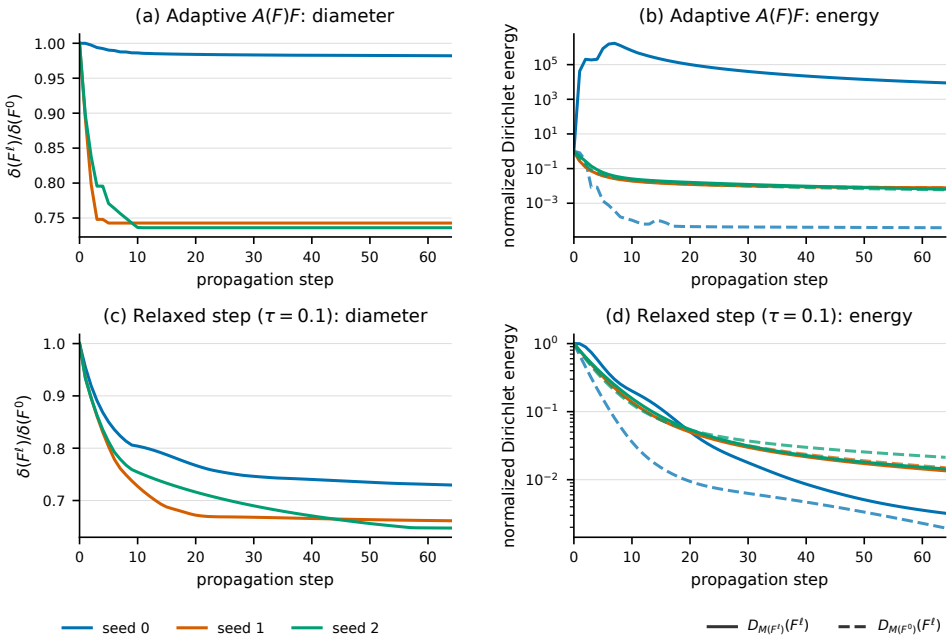

Figure 5: Dirichlet energy and diameter diagnostics initialized from the three Cora attention checkpoints. Top: softmax $A(F)F$ updates. Bottom: relaxed $(1-\tau)F + \tau A(F)F$ updates, $\tau = 0.1$. Panels (a,c) measure diameter, while (b,d) show Dirichlet energy $D_{M(F^\ell)}(F^\ell)$ (solid) and the fixed reference energy $D_{M(F^0)}(F^\ell)$ on the same states (dashed). All quantities are normalized at step zero.

Aggregation and uncertainty. Raw rows retain sequence, layer, and head identities. For layer plots, we first average heads within each sequence and then compute the mean and 1.96 standard errors across the 96 sequences. The early period is the first $\lceil L/4 \rceil$ layers and the late period starts at $\lfloor 3L/4 \rfloor + 1$. Table entries average heads and the relevant layers within a sequence before computing intervals. The one-based early/late ranges are 1–3/10–12 for both 12-layer pretrained encoders and the random BERT-base, 1–6/19–24 for BERT-large, and 1–2/5–6 for DistilBERT.

Interventions. For the same pretrained checkpoints and text blocks, we also reconstruct the per-head value tensor V. Let

$$\begin{aligned} A_{\text{full}} &= \text{softmax}_j(\log K + \zeta + 1\beta^\intercal), \\ A_{\text{met}} &= \text{softmax}_j(\log K + 1\beta^\intercal), \\ A_{\text{noK}} &= \text{softmax}_j(\zeta + 1\beta^\intercal), \\ A_{\text{no}\beta} &= \text{softmax}_j(\log K + \zeta), \\ A_{\text{sink}} &= \text{softmax}_j(1\beta^\intercal). \end{aligned}$$

We compare $A_{\text{full}}V$, $A_{\text{met}}V$, $A_{\text{noK}}V$, $A_{\text{no}\beta}V$, $A_{\text{sink}}V$, and A_{met}^2V with the vanilla output AV. Relative error is $|\widehat{O} - O|_F / |O|_F$, and cosine similarity flattens the token and head-channel dimensions. These operators are evaluated without retraining. The experiment contains 71,424 sequence–layer–head observations, with confidence intervals that use sequences as replicates.

Gauge and centering diagnostics. For affine projections $q_i = W_Q h_i + b_Q$ and $k_j = W_K h_j + b_K$, the term $q_i^\top b_K$ is constant across the key index and therefore exactly removed by row-wise softmax. We reconstruct scores both with and without b_K and record the attention/output discrepancy in float64 after float32 projection. We also count the deleted bias scalars directly from each checkpoint. This row-gauge redundancy should not be conflated with the separate change-of-basis freedom $Q \mapsto QG$, $K \mapsto KG^{-\top}$, which does not by itself reduce stored parameters without a constrained reparameterization.

To remove β, we center each head's queries over the non-padding tokens before computing scores. We numerically check that the resulting attention equals $A_{\mathrm{no}\beta}$ and that its measured β energy is zero to floating-point precision. As a negative control, we apply feature-wise LayerNorm to each projected query token; this normalizes channels but does not center over tokens.

Downstream β intervention. We use `distilbert/distilbert-base-uncased-finetuned-sst-2-english` at revision `714eb0fa89d2f80546fda750413ed43d93601a13` and the public GLUE SST-2 validation split (872 labeled examples). Inference uses eager attention, evaluation mode, native tokenization, maximum length 128, and batch size 64. For condition k, query centering is enabled in layers $1, \ldots, k$ and the checkpoint is otherwise unchanged. Key-bias removal zeros all six key-projection biases. We record predictions, logits, confidence, and negative log-likelihood per example. Reported accuracy intervals are half the distance between the 2.5th and 97.5th percentiles from 4,000 bootstrap resamples with seed 1729.

D.1 Dirichlet versus Diameter Benchmark

The Cora comparison uses the fixed public Planetoid split, seeds $\{0, 1, 2\}$, 200 epochs, 64 residual layers, hidden width 128, time step $\tau = 0.1$, one attention head, and attention ratio four. The encoder and decoder are linear, the GNN nonlinearity is the identity, attention is computed once and held constant across depth, and the feature mixer is shared across layers. The mixer uses the bounded-signed parameterization with spectral scale 0.05, temperature 0.01, and initialization standard deviation 0.1. Scaled-dot logits are divided by $\sqrt{d}$, shifted by their global maximum, and clipped to 20. Optimization uses Adam with learning rate 10^{-3} and no weight decay. Each reported test value is the value stored when validation accuracy attains its running maximum; the test set is not used for model selection.

The vanilla condition uses polynomial coefficients $(0, 1)$, i.e., $p(A) = A$. The higher-order condition changes only these coefficients to $(0, 1, 1)$, i.e., $p(A) = A + A^2$. The Laplacian condition uses graph factor $-\Delta_{M,m} = A - I$, where M is the learned symmetric conductance, $m = M\mathbf{1}$, and $A = \mathrm{diag}(m)^{-1}M$. All three graph factors enter the residual update $F^{\ell+1} = F^{\ell} + \tau p(A)F^{\ell}W$ (with $p(A) = A - I$ for the Laplacian condition). The shared learned channel mixer W is signed, and the attention scores are symmetrized before exponentiation. Parsed epoch-level learning curves are retained for every seed.

Fixed-conductance trajectory diagnostics. The training runner stores a distinct best-validation checkpoint for every variant and seed. We reload each checkpoint in evaluation mode and retain the latent state after the encoder and after every residual layer. For a state F, the diameter is computed exactly as $\max_{i,j} |\vec{f}_i - \vec{f}_j|$ using chunked pairwise distances. With the symmetric learned edge conductance M held fixed across depth, we compute $D_M(F) = \frac{1}{2}\sum_{i<j} M_{ij}|\vec{f}_i - \vec{f}_j|^2$ directly on unique undirected edges. Both quantities are normalized by their initial value.