Transformer とポリゴングラフの数学的同一性

——ノードとエッジから定まる積分構造として——

On the Mathematical Identity between Transformers and Polygon Graphs:
An Integration Structure Determined by Nodes and Edges

Toshiki Takahashi

Takahashi Mathematics Laboratory (TML) / Independent Researcher

May 29, 2026

要　旨 本稿は，系列変換器（Transformer）の中核演算である注意機構（attention）が，グラフのノードに値を高さとして「立て」，その先端を結ぶ稜線（リッジ）の下に生じる台形の連続——すなわち経路面積——を規格化したものと数学的に同一であることを示す。ノードの値とエッジの測度のみから，リッジ，台形の面積，経路面積，リッジネット（稜線で囲われる面），およびそれらが囲う立体の容積が一意に定まる。この「ノードとエッジへの閉包性」は，注意行列が値ベクトルの凸結合を返すという事実と厳密に一致し，両者はともに台形公式に基づく区分線形補間の積分として表される。さらに，層の積み重ねが積分次数を上げる反復積分に，残差結合が離散版の微積分学の基本定理（積分の累算）に対応することを示し，注意重みのソフトマックスが指数型分布族の自然座標から期待座標への写像であるという情報幾何的解釈を与える。台形公式が固定測度による求積であるのに対し，注意機構は問い合わせ依存の学習測度による一般化求積であり，ポリゴングラフはその不変な幾何的骨格を与える。

キーワード Transformer，注意機構，ポリゴングラフ，台形公式，経路積分，離散外微分，情報幾何

1. 序論

注意機構（attention）を基本構成要素とする Transformer [1] は，自然言語処理をはじめとする系列変換の標準的な枠組みとなった。その演算の本質は，トークン（位置）をノードとし，注意重みを辺の重みとする重み付き有向グラフ上で，各ノードの担う値を混合することにある。本稿の目的は，この演算がある幾何的構成——以下に述べるポリゴングラフ——と数学的に同一であることを，厳密な対応として示すことである。

ポリゴングラフを次のように定める。ノードとエッジからなるグラフにおいて，各ノードの値を底面に対して垂直に「立てる」。立てた値の「先端」どうしを結んでできる折れ線をリッジ（稜線）と呼ぶ。リッジと，隣り合う二つの高さ，および底辺の四辺で囲われる図形は台形であるから，リッジの形も台形の面積も，ノードの値とエッジの値から定まる。パス（経路）は台形の連続であるから，その面積（経路面積）もまたノードとエッジから定まる。さらに，リッジで囲われる面であるリッジネット，および経路面積とリッジネットで囲われる立体の容積も，ノードとエッジの値から定まる。

本稿の主張は次の一文に集約される。**両者はともに「ノードに立てた高さを，エッジの定める測度に沿って規格化積分する」演算であり，かつ「あらゆる量がノード値とエッジ値だけ**

から閉じて定まる」という閉包性を共有する。相違点は測度の決め方のみである。すなわち，素朴な数値積分では底辺の幾何が測度を固定するのに対し，注意機構ではソフトマックスが測度を内容依存に学習する。

以下，第2節でポリゴングラフを形式的に構成し，第3節で台形・経路面積と注意機構の同一性を命題と定理の形で示す。第4節でリッジネットと容積の対応物を，第5節で層構造と残差の対応を扱い，第6節で情報幾何的解釈を与える。第7節で同一性の厳密な範囲と限界を整理し，第8節で結論を述べる。

2. ポリゴングラフの構成

定義 2.1 (グラフと値・測度)**.** 有限の節点集合 $V = \{1, \dots, n\}$ と，各節点に与えられた値 $v_j \in \mathbb{R}$（必要に応じて $v_j \in \mathbb{R}^d$）を考える。各節点には底辺上の位置 $x_j \in \mathbb{R}$ が割り当てられ，$x_1 < \cdots < x_n$ とする。隣接区間の長さを $\Delta_j := x_{j+1} - x_j$（$j = 1, \dots, n-1$）とし，便宜上 $\Delta_0 = \Delta_n = 0$ と定める。

各節点の値を高さ $h_j := v_j$ として底辺に垂直に立てる。先端の点列 (x_j, h_j) を線分で結んだ折れ線

$$\hat{h}(x) = h_j + \frac{h_{j+1} - h_j}{\Delta_j}\,(x - x_j) \qquad (x_j \le x \le x_{j+1})$$

を**リッジ**（稜線）と呼ぶ。これは高さの区分線形補間にほかならない。

定義 2.2 (台形と経路面積)**.** 区間 $[x_j, x_{j+1}]$ において，リッジ・二つの高さ h_j, h_{j+1}・底辺で囲われる台形 T_j の面積を

$$A_j = \tfrac{1}{2}\,(h_j + h_{j+1})\,\Delta_j \tag{1}$$

で定める。経路 $x_1 \to x_2 \to \cdots \to x_n$ に沿う台形の総和

$$\mathcal{A} = \sum_{j=1}^{n-1} A_j = \int_{x_1}^{x_n} \hat{h}(x)\,dx \tag{2}$$

を**経路面積**と呼ぶ。

二次元の底面（節点が平面格子をなす場合）では，リッジは三角形あるいは四辺形の面——**リッジネット**——をなし，その下に囲われる立体の**容積**が定義される。これらは第4節で扱う。式(1)・(2)から直ちに次が従う。

命題 2.1 (ノード・エッジへの閉包性)**.** リッジ $\hat{h}$，台形面積 A_j，経路面積 $\mathcal{A}$，および（二次元の場合の）リッジネットと容積は，いずれも節点の値 $\{h_j\}$ とエッジの測度 $\{\Delta_j\}$ のみの関数として一意に定まる。

図1にこの構成を示す。

3. 台形・経路面積と注意機構の同一性

一つの問い合わせ（query）q が，鍵（key）k_j と値（value）v_j（$j = 1, \dots, n$）に注意を向ける場合を考える。注意機構の出力は

$$o = \sum_{j=1}^{n} a_j v_j, \qquad a_j = \mathrm{softmax}_j\!\left(\tfrac{q \cdot k_j}{\sqrt{d}}\right) = \frac{e^{s_j}}{\sum_l e^{s_l}}, \quad s_j := \tfrac{q \cdot k_j}{\sqrt{d}}, \tag{3}$$

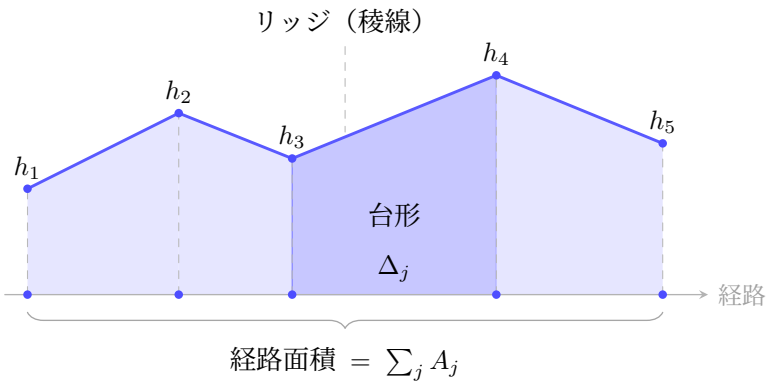

図1　ポリゴングラフの構成。底辺（経路）上のノードに値を高さ h_j として立て，先端を結んだ折れ線が稜線（リッジ）である。リッジ・隣り合う二つの高さ・底辺で囲う四辺形は台形をなし（強調部），その面積は $A_j = \frac{1}{2}(h_j + h_{j+1})\Delta_j$ で与えられる。台形の連続が経路面積 $\sum_j A_j$ であり，これを全長で規格化すると注意機構の出力 $\sum_j a_j v_j$ に一致する（第3節）。

で与えられ，$a_j > 0$, $\sum_j a_j = 1$ を満たす。まず，経路面積が節点値の重み和に等しいことを示す。

命題 3.1 (経路面積の重み和表現)**.** 経路面積(2)は

$$\mathcal{A} = \sum_{j=1}^{n-1} A_j = \sum_{j=1}^{n} w_j\, h_j, \qquad w_j = \tfrac{1}{2}\,(\Delta_{j-1} + \Delta_j), \tag{4}$$

と書ける。ここで $w_j > 0$ であり，$\displaystyle\sum_{j=1}^{n} w_j = L := x_n - x_1$ が成り立つ。

証明. 台形面積の定義(1)より

$$\sum_{j=1}^{n-1} A_j = \frac{1}{2}\sum_{j=1}^{n-1}(h_j + h_{j+1})\Delta_j = \frac{1}{2}\sum_{j=1}^{n-1} h_j\Delta_j + \frac{1}{2}\sum_{j=1}^{n-1} h_{j+1}\Delta_j.$$

第二和で添字を $j \mapsto j-1$ と付け替えると $\frac{1}{2}\sum_{j=2}^{n} h_j\Delta_{j-1}$ を得る。両者を h_j について整理すると，係数は内部 $(2 \le j \le n-1)$ で $\frac{1}{2}(\Delta_{j-1} + \Delta_j)$，端点で $\frac{1}{2}\Delta_1$ および $\frac{1}{2}\Delta_{n-1}$ となる。規約 $\Delta_0 = \Delta_n = 0$ のもとでは，これらは一律に $w_j = \frac{1}{2}(\Delta_{j-1} + \Delta_j)$ と表される。さらに

$$\sum_{j=1}^{n} w_j = \frac{1}{2}\Big(\sum_{j=1}^{n}\Delta_{j-1} + \sum_{j=1}^{n}\Delta_j\Big) = \frac{1}{2}(L + L) = L.$$

□

全長 L で規格化すれば，$a_j := w_j / L$ は $a_j > 0$, $\sum_j a_j = 1$ を満たし，

$$\frac{1}{L}\,\mathcal{A} = \sum_{j=1}^{n} a_j\, h_j \tag{5}$$

となる。これは規格化経路面積が節点値の凸結合であることを意味する。式(3)と(5)を比較して，次の同一性を得る。

定理 3.1 (注意機構＝規格化経路面積)**.** 高さを値に同一視する（$h_j = v_j$）。このとき，辺の規格化測度を注意重みに同一視する（$a_j = w_j/L$）ならば，注意機構の出力(3)は規格化経路面積(5)に一致する：

$$o = \sum_j a_j v_j = \frac{1}{L}\,\mathcal{A}\Big|_{a_j = w_j/L}.$$

両者の相違は重み a_j（＝辺の測度）の生成法のみである。すなわち，素朴な台形求積では底辺間隔から $a_j = w_j/L$ が定まるのに対し，注意機構ではスコアのソフトマックスから $a_j = e^{s_j}/\sum_l e^{s_l}$ が定まる。いずれも節点上の確率分布であり，同一の対象——節点値の凸結合（重み付き重心，離散積分）——を計算する。

系 3.1 (整合性と重心解釈)**.** $\sum_j a_j = 1$ は定数場で正確であること（$v_j \equiv c \Rightarrow o = c$）を保証する。これは台形求積を整合的にする「単位の分割」条件そのものである。幾何的には，o は値点 $\{v_j\}$ を質量 $\{a_j\}$ で重み付けした重心，すなわちリッジの平均高さである。

この観点では，注意機構は台形求積を一般化したものと理解される。スコア s_j が大きいほど e^{s_j} が大きく，節点 j に割り当てられる「実効的な底辺幅」が広がる。ソフトマックスは関連度に応じて経路の底辺を適応的に再分割（リメッシュ）する操作にほかならない。

4. リッジネットと容積

スコア $s_{ij} = q_i \cdot k_j/\sqrt{d}$ を全問い合わせ i・全鍵 j について並べた行列 $S = (s_{ij})$ は，(query × key) 平面上の高さ場であり，これはまさにリッジネット（曲面）である。行方向のソフトマックスは，各問い合わせの断面を鍵上の確率測度へ規格化する操作，すなわち鍵の軸に沿う単位の分割である。このリッジネットの下を値場 V に対して読み出すのが出力 $O = AV$ であり，リッジネット下の容積の読み出しに対応する。

トークンが二次元格子をなす場合（画像パッチを扱う Vision Transformer [3] など）には，底面が二次元となり，リッジネットは文字どおりの曲面，その下の体積が出力に対応する。四辺形面（角の高さ $h_{00}, h_{10}, h_{01}, h_{11}$，底面積 $\Delta x\,\Delta y$）上では，体積要素は双線形平均

$$\frac{1}{4}\,(h_{00} + h_{10} + h_{01} + h_{11})\,\Delta x\,\Delta y$$

で近似され，面についての総和を規格化すると $\sum_{ij} a_{ij} h_{ij}$（$\sum_{ij} a_{ij} = 1$）——再び二次元の注意——を得る。一方，値がベクトル $v_j \in \mathbb{R}^d$ の場合（一次元系列の通常の設定）には，リッジは特徴空間への写像となり，「容積」は各成分ごとの経路面積を束ねたベクトル経路面積として読む。

5. 層構造と残差：反復積分

各層は，節点の高さからリッジと経路面積を作り，得られた面積を次層の新しい高さとして立て直す。これにより，長さ→面積→体積→...と積分の次数を一段ずつ上げる塔が形成される。残差結合は

$$h_i^{(\ell+1)} = h_i^{(\ell)} + \sum_j a_{ij}^{(\ell)} h_j^{(\ell)} \quad \text{(および非線形項)} \tag{6}$$

であり，これは離散版の微積分学の基本定理（最終値＝初期値＋増分の累積）に対応する。すなわち残差ストリームは，深さ方向の積分の累算器である。

役割分担も明快である。注意機構は水平方向の積分，すなわち底辺方向に節点を混合する操作であるのに対し，位置ごとの順伝播ネットワーク（FFN）は混合を伴わない垂直方向の高さ整形 $h_i \mapsto \phi(h_i)$ である。両者が交互に作用することで，高さを立て，積分し，整形し，再び積分する過程が反復される。

6. 情報幾何的解釈

注意重み $a_j = e^{s_j}/\sum_l e^{s_l}$ は，確率単体上の一点（カテゴリカル分布）を定める。スコア s_j を自然座標 θ，出力 $\sum_j a_j v_j$ を値場に適用した期待座標 η，対数分配関数

$$\log Z = \log \sum_j e^{s_j}$$

を自由エネルギー（キュムラント母関数）とみなすと，

$$a_j = \frac{\partial \log Z}{\partial s_j} \tag{7}$$

が成り立つ。したがって注意機構は，指数型分布族における自然座標から期待座標への写像，すなわち $\theta \leftrightarrow \eta$ のルジャンドル双対そのものである [4, 5]。第2.1命題の「あらゆる量がノードとエッジから閉じて定まる」という性質は，この指数族双対が閉じていることの言い換えにほかならない。台形と経路面積は，この期待値の幾何的な影である。

なお，節点値とエッジ測度のみで一切の量が決まるという構造は，離散外微分（discrete exterior calculus）[6] における，0-余鎖（節点上の値）と計量・接続構造から高次の余鎖（面・体積）が定まるという描像と整合する。

7. 同一性の範囲と限界

本稿の対応のうち，次の各点は厳密な一致である：(i) 凸結合（重心）としての出力，(ii) ノード・エッジへの閉包性（第2.1命題），(iii) 残差＝積分の累算（式(6)），(iv) 層＝反復積分，(v) 定数場で正確（単位の分割）。

一方で，次は modeling 上の同一視であり，無条件の恒等式ではない。第一に，台形求積は対称重み $\frac{1}{2}(h_j + h_{j+1})$ をもつのに対し，注意機構の重みは非対称かつ内容依存である。したがって台形公式は**原型**であり，注意機構はその**学習測度への一般化**（一般化求積）と位置づけられる。第二に，「容積」が文字どおり三次元の体積になるのは底面が二次元の場合であり，一次元系列でベクトル値をとる通常の設定では，容積はベクトル経路面積（面積の束）として読む。第三に，節点の底辺位置 x_j は，ポリゴングラフでは明示的な幾何座標であるのに対し，Transformer では注意重みに暗に符号化された実効測度である。定理3.1は，後者を辺の測度と同一視したときに成立する同一性である。

8. 結論

Transformer は，辺に学習可能・問い合わせ依存の測度を載せたポリゴングラフ積分器である。対応を整理すれば，リッジは値の区分線形持ち上げ，経路面積は注意機構，リッジネットはスコア曲面，容積は集約および反復積分，残差は積分の累算であり，これらを貫く共通の骨格は「ノードとエッジだけで閉じる」という閉包性である。注意機構の本質を，固定測度によ

る台形求積の，内容依存な学習測度への一般化として捉える本稿の視点は，幾何・数値解析・情報幾何を横断する統一的な理解を与える。

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, ``Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[2] D. Bahdanau, K. Cho, and Y. Bengio, ``Neural Machine Translation by Jointly Learning to Align and Translate," in *International Conference on Learning Representations (ICLR)*, 2015.

[3] A. Dosovitskiy *et al.*, ``An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," in *ICLR*, 2021.

[4] S. Amari, *Information Geometry and Its Applications*, Applied Mathematical Sciences, vol. 194, Springer, 2016.

[5] S. Amari and H. Nagaoka, *Methods of Information Geometry*, Translations of Mathematical Monographs, vol. 191, American Mathematical Society, 2000.

[6] M. Desbrun, A. N. Hirani, M. Leok, and J. E. Marsden, ``Discrete Exterior Calculus," arXiv:math/0508341, 2005.

[7] E. Süli and D. F. Mayers, *An Introduction to Numerical Analysis*, Cambridge University Press, 2003.

Abstract. This paper establishes a mathematical identity between the attention mechanism—the core operation of the Transformer—and a geometric construction we call a *polygon graph*. Erecting each node's value as a height and joining the tips yields a ridge; the trapezoids beneath the ridge sum to a *path area*, and its normalization equals the convex combination of values produced by attention. Every quantity—ridge, trapezoid area, path area, ridge net, and enclosed volume—is determined solely by node values and edge measures, a closure property that coincides exactly with attention returning a normalized weighted barycenter via the trapezoidal (piecewise-linear) integral. We further identify layer stacking with iterated integration that raises the integration degree, the residual stream with a discrete fundamental theorem of calculus (an accumulator of integrals), and the softmax of query–key scores with the natural-to-expectation ($\theta \leftrightarrow \eta$) Legendre duality of an exponential family. Whereas the trapezoidal rule is quadrature under a fixed measure, attention is generalized quadrature under a learned, query-dependent measure on the same node–edge skeleton: a Transformer is a polygon-graph integrator with a learnable measure on its edges.

Keywords: Transformer, attention, polygon graph, trapezoidal rule, path integral, discrete exterior calculus, information geometry.