基盤モデルの作る潜在空間の性質

早稲田大学 理工学術院 村田 昇，赤塚 育海

Noboru Murata, Ikumi Akatsuka
Waseda University, Faculty of Science and Engineering

近年の深層学習の進展に伴い，事前学習された基盤モデルが画像処理や自然言語処理において中心的な役割を担っている．本稿では，CNN や Transformer など情報処理機構の異なるモデルが共通して持つ潜在変数の重み付き和という情報の集約構造に着目し，モデルの出力ベクトルを指数型分布族の期待値パラメタとして再解釈する．この対応付けに基づき，情報幾何の枠組みを用いて潜在空間を統計多様体として定式化し，重み付きの潜在表現による標本統計量から計量，接続，スカラー曲率を直接推定する手法を提案する．実験的検証により，潜在空間が学習問題により異なる曲率をもつこと，Fisher 計量に基づく測地距離がユークリッド距離に代わって意味的に滑らかな遷移を実現すること，モデル間の幾何構造が学習過程や学習データの系譜を強く反映することが示唆された．本稿で述べる方法は，モデル評価の視点を単に出力や性能から評価するのではなく，学習器内部に獲得される情報処理のための空間の幾何構造として捉えることを提案するものであり，基盤モデルの解釈性向上や次世代のモデルの設計の新たな指針となることを目的とする．

1 はじめに

計算機環境の発展にともない，画像処理や自然言語処理などの分野では大規模なデータセットの利用が可能となり，高精度なアノテーションの付加とデータの整理が進んだ結果，CNN や Transformer などに基づいた多様な情報処理機構が提案され，画像や言語データに対する特徴抽出能力が飛躍的に向上した．特に，事前学習された強力な基盤モデル (foundation model) が公開され，提供される汎用的なベクトル表現を用いて多様な課題への応用が可能となった．

基盤モデルの働きの一つは，入力 x を高次元ベクトル空間に写す写像 $\eta(x)$ を提供することである．この写像 $\eta(x)$ は，画像や言語の意味的な特徴を捉えたものであり，オブジェクト埋め込み (object embedding) と呼ばれる．この写像により得られる値は，分類，検出，セグメンテーション，生成など多様な課題に再利用可能であり，入力対象の汎用的な特徴量としての役割を果たす．

基盤モデルの一般的な構造は以下のように記述される．画像，音声，文書などの入力データを x と書くことにし，モデルの出力は $\eta(x)$ で表す．入力 x は通常適切な方法でトークンと呼ばれる要素に分解されるので，これを $t_{\lambda'}(x),\ \lambda' \in \Lambda'$ と書くことにする．Λ' は例えば画像や文書の内での要素の位置（location）に対応する添字集合である．出力の計算過程で現れる潜在変数を $z_\lambda(x),\ \lambda \in \Lambda$ と書く．Λ は同様に潜在変数の位置に対応する添字集合である．多くの基盤モデルの情報変換の流れは，以下の構造を持つ：

$$x \mapsto T(x) = \{t_{\lambda'}(x),\ \lambda' \in \Lambda'\} \mapsto Z(x) = \{z_\lambda(x),\ \lambda \in \Lambda\} \mapsto \eta(x) = \sum_{\lambda \in \Lambda} w_\lambda z_\lambda(x).$$

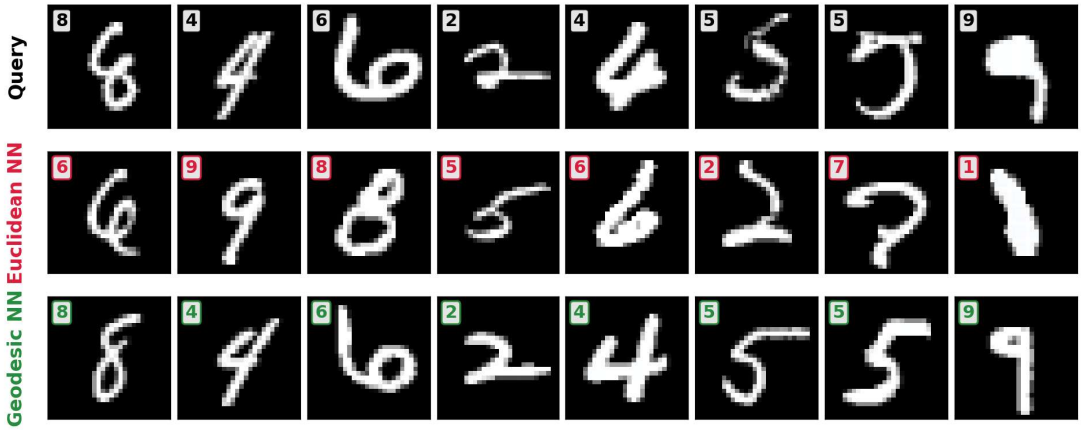

Figure 1: ユークリッド距離と測地距離にもとづく近傍点の違い．対象の数字 (上段)，ユークリッド距離 (中段)，測地距離 (下段)．ユークリッド距離では異なる数字となるが，測地距離では同じ数字が最近傍となっている．

ただし重み w_λ は $w_\lambda \geq 0$, $\sum_{\lambda\in\Lambda} w_\lambda = 1$ を満たす．重み付き潜在変数の和により入力対象 x からベクトル表現 $\eta(x)$ を得る情報処理は，例えば

- CNN の Global Average Pooling（GAP）（例：LeCun et al. 1998 [1] ）
- Transformer の Attention 機構（例：Vaswani et al. 2017 [2] , Dosovitskiy et al. 2020 [3] ）

に見られ，次節で具体的な学習器の構造を述べる．しかし，こうして得られたベクトル表現 $\eta(x) \in \mathcal{M} \subset \mathbb{R}^d$ において，単純なユークリッド距離

$$d_{\mathrm{euc}}(\eta(x_1), \eta(x_2)) = \|\eta(x_1) - \eta(x_2)\|_2$$

で意味的な類似度を測ろうとすると，問題が生じることがある．Figure 1 は MNIST データ (0〜9 の手書き数字の画像データセット) で判別問題を学習した CNN の潜在空間において，ユークリッド距離では正しく類似性を捉えられない例を示したものである．上段は対象の数字，中段はユークリッド距離で最も近いデータ，下段は後に定義する計量 g_{ij} を考慮した測地距離

$$d_{\mathrm{geo}}(\eta(x_1), \eta(x_2)) = \inf_{\gamma} \int_0^1 \sqrt{g^{ij}(\gamma(t))\, \dot{\gamma}_i(t) \dot{\gamma}_j(t)}\, dt$$

で最も近いデータを示している．ここで $\gamma : [0,1] \to \mathcal{M}$ は $\eta(x_1)$ から $\eta(x_2)$ へ滑らかな曲線を表す．数値計算上は，近傍点同士は 2 点の計量の平均を用いた 2 次近似で距離を計算して疎なグラフを作成し，近傍点以外は Isomap [4] と同様な考え方でグラフ上の最短経路問題を解くことによって測地距離の近似を行っている．この例は，モデルが構築する潜在空間が平坦なユークリッド空間ではなく，データの意味的構造に従って複雑に湾曲していることを示唆している．

次節では，出力 $\eta(x)$ が潜在変数 $z_\lambda(x)$ の重み付き平均とみなせることを利用して，基盤モデルを確率分布と対応づけ，その幾何学的性質を考察することを考える．

2 基盤モデルの例

前節で導出した基盤モデルに共通な集約構造を踏まえ，本節では代表的な三つの機構 (CNN，Transformer，ViT) における具体的な計算過程を整理する．潜在変数の平均として得られたベクトル表現を確率分布の期待値パラメタと対応付ける理論的枠組みをまとめ，情報幾何的な解析へのつながりを示す．

2.1 畳み込みニューラルネットワーク (CNN) と大域的平均プーリング

CNN [1] は，画像などの空間データに対して局所的な畳み込みとプーリングを繰り返し適用し，階層的な特徴量マップ (抽出された特徴量を画像と対応するように空間的に配置したもの) を生成する．畳み込みは，対象となる画素の近傍の特徴量の重み付き和と非線形変換を，平行移動させながら繰り返し適用する操作であり，プーリングは，局所的な領域の平均などの統計量を取り出すことで画素数を削減 (ダウンサンプリング) し，データの圧縮と位置ズレへの耐性を高める操作である．このため，入力画像 x に対して，トークン $T(x) = \{t_{\lambda'}(x),\ \lambda' \in \Lambda'\}$ の添字集合 Λ' は画素の位置と考えれば良いが，潜在空間 (特徴量マップ) の添字集合 Λ の大きさは一般に画素数とは異なる．最終的なベクトル表現を得る過程では，大域的平均プーリング（Global Average Pooling; GAP）が広く用いられる．GAP は，各特徴量マップに含まれる全画素の平均を計算することで，位置情報を圧縮し統計的に集約された表現を得る．数学的には，位置添字集合 $\Lambda = \{\lambda = (i, j) \mid 1 \le i \le H,\ 1 \le j \le W\}$ における潜在変数 $z_\lambda(x)$ に対して一律の重み $w_\lambda = 1/(HW) = 1/|\Lambda|$ を割り当て，

$$\eta(x) = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} z_{(i,j)}(x) = \sum_{\lambda \in \Lambda} \frac{1}{|\Lambda|} z_\lambda(x)$$

の形で集約される．この場合，重みは位置に依存せず固定されており，画像内の「局所的な特徴がどれだけ含まれているか」という量のみを保持する．

2.2 Transformer と自己注意機構

Transformer 機構 [2] に基づく大規模言語モデル (Large Language Model; LLM) では，入力文を語句や形態素などのトークン列として分解し，自己注意機構（Self-Attention）を用いて文脈に応じた情報の集約を行う．一般に Transformer は複数の層からなり，各層ではベクトル集合 $Z = \{z_\lambda,\ \lambda \in \Lambda\}$ をベクトル集合 $Z' = \{z'_\lambda,\ \lambda \in \Lambda\}$ に変換する処理が自己注意機構を利用して実現される．さまざまな拡張があるが，基本となる情報変換は以下のようにまとめられる．ベクトル集合 $Z = \{z_\lambda,\ \lambda \in \Lambda\}$ から Value $V = \{v_\lambda\}$, Key $K = \{k_\lambda\}$, Query $Q = \{q_\lambda\}$ の 3 種類のベクトル集合が計算される．位置 μ のトークンに対する注意重み (Attention Weight) $w^\mu_\lambda,\ \lambda \in \Lambda$ は，Query q_μ と Key k_λ の内積に対して SoftMax 関数を適用することで動的に計算される:

$$w^\mu_\lambda = \frac{\exp(c\langle q_\mu, k_\lambda \rangle)}{\sum_{\nu \in \Lambda} \exp(c\langle q_\mu, k_\nu \rangle)} = \mathrm{SoftMax}(q_\mu, K).$$

変換後のベクトル集合 $Z' = \{z'_\mu\ \mu \in \Lambda\}$ はこの重みを用いた Value $V = \{v_\lambda,\ \lambda \in \Lambda\}$ の加重平均として得られる:

$$z'_\mu = \sum_{\lambda \in \Lambda} w^\mu_\lambda v_\lambda.$$

通常は最終層における特定のトークンの表現が入力のベクトル表現として用いられるため，潜在変数にあたる Value ベクトル $z_\lambda(x)$ のこのトークンに対する注意重み w_λ を用いた加重平均として表現される:

$$\eta(x) = \sum_{\lambda \in \Lambda} w_\lambda z_\lambda(x).$$

CNN とは異なり，重みは文脈に依存して変化するため，より柔軟で文脈に依存した特徴量の集約が実現される．

2.3 Vision Transformer (ViT) とパッチ列の集約

Vision Transformer (ViT) は，画像を固定サイズのパッチに分割し，それをトークン列として扱うことで，画像処理に Transformer の枠組みを適用したモデルである．画像全体のベクトル表現 $\eta(x)$ を得る方法としては，特殊なトークンを用いた注意機構による集約，または最終層のパッチ表現に対する大域的平均プーリングが主に用いられる．いずれの場合も，パッチ位置 $\lambda \in \Lambda$ における潜在変数 $z_\lambda(x)$ に対して重み w_λ（注意重みまたは一様な $1/|\Lambda|$）を適用し，

$$\eta(x) = \sum_{\lambda \in \Lambda} w_\lambda z_\lambda(x)$$

の重み付き平均でベクトル表現を構成する．ViT はパッチ内の局所的処理と，Transformer 特有の大域的な依存関係を捉える処理を合わせ持ち，より複雑な潜在空間の幾何構造を獲得する．

2.4 確率分布への対応付け

上記三つのモデルは情報処理の機構こそ異なるが，いずれも出力 $\eta(x)$ を潜在変数 $\{z_\lambda(x),\ \lambda \in \Lambda\}$ の重み付き和として記述している．したがって基盤モデルの出力 $\eta(x)$ は，潜在空間の重み付きデータ集合 $\{(z_\lambda(x), w_\lambda)\}$ から計算される標本平均 $\mathbb{E}_{Z \sim \mathcal{D}_x}[Z] = \sum_{\lambda \in \Lambda} w_\lambda z_\lambda(x)$ に対応し，入力 x を分布族 $\mathcal{M}$ 内の一点（期待値パラメタ）へ写す写像と解釈できる．

この対応付けは，潜在空間の解析において新たな視点を与える．まず，特徴量を単なるベクトル空間上の点ではなく，確率分布そのもの，あるいは分布の期待値パラメタとみなすことで，基盤モデル同士をそれらが構成する確率分布族として比較することが可能となる．次節で述べる情報幾何による統計多様体の理論では，Fisher 情報量にもとづく Fisher 計量が分布間の類似度を測る自然な計量として定義される．更にこの枠組みは基盤モデルが構築する潜在空間の幾何学的特徴を定量化することを可能にする．計量や接続係数，スカラー曲率などの幾何学的な量は，モデルがデータをどのように階層化し，どのように曲がった多様体上の点として表現しているかを特徴付ける．例えば，負の曲率領域の存在は双曲的な情報の埋め込みを示唆し，モデルの汎化能力や表現の多様性と関係する可能性がある．次節では，指数型分布族への対応付けを考え，幾何学量の具体的な計算方法を述べる．

3 指数型分布族の情報幾何

3.1 統計多様体の導入

先に述べたように，前節で導出した基盤モデルに共通な集約構造

$$\eta(x) = \sum_{\lambda \in \Lambda} w_\lambda z_\lambda(x)$$

は，統計学における期待値（標本平均）の定義と形式的に一致する．この性質に着目し，モデルの出力 $\eta(x)$ を単なるベクトル空間上の点ではなく，指数型分布族 $\mathcal{M}$ における期待値パラメタとして捉え直す．潜在変数 Z を測度空間 $(\mathcal{Z}, \mu)$ 上の確率変数とし，指数型分布族を

$$\mathcal{M} = \left\{ P_\theta,\ \theta \in \Theta \,\middle|\, p(z;\theta) = \exp\bigl(c(z) + \langle \theta, z \rangle - \psi(\theta)\bigr) \right\}$$

で定義する．ここで $\psi(\theta)$ はポテンシャル関数，θ は正準パラメタである．正規化条件 $\int_{\mathcal{Z}} p(z;\theta) d\mu(z) = 1$ を θ で微分すると

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta^i} \int_{\mathcal{Z}} p(z;\theta) d\mu(z) = \int_{\mathcal{Z}} p(z;\theta) \frac{\partial}{\partial \theta^i} \log p(z;\theta) d\mu(z) \\ &= \mathbb{E}_{Z \sim P_\theta}[Z_i - \partial_i \psi(\theta)] \end{aligned}$$

であることから，期待値パラメタは $\eta_i = \partial_i \psi(\theta)$ で与えられる．適当な条件のもとで $\partial\psi$ は全単射となるため，平均 η も分布を一意に定める．基盤モデルの出力 $\eta(x)$ は重み付きデータ集合 $\mathcal{D}_x = \{(z_\lambda(x), w_\lambda)\}_{\lambda\in\Lambda}$ の標本平均 $\mathbb{E}_{Z\sim\mathcal{D}_x}[Z] = \sum_{\lambda\in\Lambda} w_\lambda z_\lambda$ に対応し，入力 x を分布族 $\mathcal{M}$ 内の一点へ写す写像と解釈できる．この対応付けは，構造の異なるモデルの潜在空間の違いを確率分布族の違いとして捉え，ベクトル表現の背後にある分布の広がり (分散) や空間の湾曲 (曲率) といった高次の幾何情報を用いた比較を可能にすることが期待される．

3.2 情報幾何学的量の定義と幾何学的意義

情報幾何 [5] は微分幾何を用いて確率分布の空間を扱う方法論で，Fisher 情報量に基づく計量と，内積を保存する平行移動を定義する双対接続が特徴である．分布族 $\mathcal{M}$ 上の一点 θ における接空間 $T_\theta\mathcal{M}$ の基底を

$$\partial_i \leftrightarrow \partial_i \log p(z;\theta) = \frac{\partial}{\partial\theta_i}\log p(z;\theta) \in T_\theta\mathcal{M}$$

で与える．Riemann 計量は Fisher 情報量で定義され，その局所座標表示は

$$\begin{aligned} g_{ij}(\theta) &= \langle \partial_i,\ \partial_j \rangle \\ &= \mathbb{E}_{Z\sim P_\theta}[\partial_i \log p(Z;\theta)\partial_j \log p(Z;\theta)] \\ &= \mathbb{E}_{Z\sim P_\theta}[(Z_i-\eta_i)(Z_j-\eta_j)] = \partial_i\partial_j\psi(\theta) \end{aligned}$$

と書ける．計量 g_{ij} は潜在変数の分散で計算され，潜在空間の各点におけるいわば「縮尺」を決定する．直感的にはパラメタ θ の推定のしやすさを表している．

アファイン接続の局所座標表示 (接続係数) は $\nabla_{\partial_i}\partial_j = \sum_k \Gamma_{ij}{}^k \partial_k$ で与えられるが，情報幾何では α-接続

$$\begin{aligned} \Gamma^{(\alpha)}_{ijk}(\theta) &= \mathbb{E}_{Z\sim P_\theta}\left[\left(\partial_i\partial_j \log p_\theta(Z) + \frac{1-\alpha}{2}\partial_i \log p_\theta(Z)\partial_j \log p_\theta(Z)\right)\partial_k \log p_\theta(Z)\right] \\ &= \frac{1-\alpha}{2}\,\mathbb{E}_{Z\sim P_\theta}\left[(Z_i-\eta_i)(Z_j-\eta_j)(Z_k-\eta_k)\right] \\ \Gamma^{(\alpha)k}_{ij} &= \sum_h g^{hk}\Gamma^{(\alpha)}_{ijh} \end{aligned}$$

を考える．双対な 2 つの接続（α-接続と $-\alpha$-接続）でそれぞれ平行移動した 2 つのベクトルは内積を保存する．接続は空間内でベクトルをどのように平行に移動させるかを定めるルールであり，分布の構造に沿った自然な経路 (測地線) を与えるが，潜在変数の三次モーメント (歪度) に依存することがわかる．

曲率に関する量として曲率テンソル，Ricci テンソル，およびスカラー曲率を定義する．それぞれの局所座標表示は

$$\begin{aligned} R_{ijkm} &= (\partial_i\Gamma_{jk}{}^s - \partial_j\Gamma_{ik}{}^s)g_{sm} + (\Gamma_{irm}\Gamma_{jk}{}^r - \Gamma_{jrm}\Gamma_{ik}{}^r) \\ R_{ij} &= \sum_{km} R_{kijm}g^{km} \\ \kappa &= \frac{1}{n(n-1)}R_{ijkm}g^{im}g^{jk} = \frac{1}{n(n-1)}\sum_{ij} R_{ij}g^{jk} \end{aligned}$$

で与えられる．スカラー曲率 κ は空間の局所的な形状を単一の数値で特徴付ける．$\kappa > 0$ は球面的（高密度凝集），$\kappa = 0$ は平坦，$\kappa < 0$ は双曲面的（階層的・ツリー状構造）をそれぞれ示唆し，負の曲率は階層的な情報構造の効率的な埋め込みを反映していると考えられる．

3.3 標本統計量に基づく幾何学量の推定

基盤モデルが構成する分布族 $\mathcal{M}$ において，$c(z)$ と $\psi(\theta)$ の具体形は未知であり，潜在変数 z の高次元性やデータ数の制約から密度関数の直接推定は困難である．したがって，上記の幾何学量を重み付き標本平均 $\mathbb{E}_{Z\sim\mathcal{D}_x}[\cdot]$ を用いて推定する．

まず，期待値パラメタの第 i 成分は，標本平均を用いて

$$\hat{\eta}_i = \mathbb{E}_{Z\sim\mathcal{D}_x}[Z_i]$$

で得られる．Fisher 計量は標本分散を用いて

$$\hat{g}_{ij} = \mathbb{E}_{Z\sim\mathcal{D}_x}[(Z_i-\hat{\eta}_i)(Z_j-\hat{\eta}_j)]$$

と推定される．これは中心化した観測行列の積で書けるため計算は容易である．

α-接続係数の推定量は標本歪度 (3 次モーメント) より

$$\hat{\Gamma}^{(\alpha)}_{ijk} = \frac{1-\alpha}{2}\,\mathbb{E}_{Z\sim\mathcal{D}_x}\big[(Z_i-\hat{\eta}_i)(Z_j-\hat{\eta}_j)(Z_k-\hat{\eta}_k)\big]$$

で与えられる．接続係数 $\hat{\Gamma}^{(\alpha)k}_{ij}$ を得るには，前述の計量の推定量の逆行列 $\hat{g}^{hk}$ を用いる必要があるが，高次元空間では共分散行列が非正則となる場合があるため，数値計算上は，一般化逆行列の採用，Tikhonov 正則化 $V+\lambda I$，または Woodbury の恒等式による低ランク近似などを導入し，数値的安定性を確保する必要がある．

曲率テンソルおよびスカラー曲率も同様に，経験分布から算出可能である．接続係数の微分を計算するためには，計量の逆行列の微分などの計算が必要となり，記述が繁雑となるので，ここでは割愛する．

以上のように，モデル内部から抽出可能な中間表現と重みを標本とみなすことにより，必要な幾何学的量を順次推定することができる．

4 実験的検証

本章では、前節で定式化した情報幾何学的枠組みを実際の基盤モデルに適用し、その有効性を検証する [6]．実験は「個別モデルの幾何学的性質の抽出」「測地距離による意味的類似度の再現」「複数モデル間の概念空間比較」の三つの段階に分けて実施し，提案手法が潜在空間の構造を定量的に記述できることを実証する．

4.1 潜在空間の曲率解析

基盤モデルが構築する潜在空間の局所的・大域的な形状を把握するため，入力データごとのスカラー曲率 κ を算出した．対象モデルとして GPT-2-Large [7] を用いた．入力データには HuggingFace `datasets` ライブラリ [8] を介して，Wikipedia, WikiText [9] IMDB [10] AG-News, Yelp, Amazon, DBpedia [11] Tweet, RottenTomatoes の 10 種類のコーパスから 10 文書ずつ抽出した．各文書 x に対応する期待値パラメタ $\eta(x)$ において，第 3 節で定義した Fisher 計量 $\hat{g}_{ij}$ と接続係数 $\hat{\Gamma}_{ijk}$ からスカラー曲率 $\kappa(x)$ を逐次推定した．なお，スカラー曲率 κ は，計算量問題のため，期待値座標 η の PCA 上位 2 主成分が張る部分多様体上で評価した．

Figure 2 は，推定された Fisher 計量に基づく近傍グラフ上の最短経路距離を用いた Isomap [4] による 2 次元埋め込みである．各点の色はスカラー曲率 κ の大きさを表しており，空間全体が明確に負の領域（$\kappa<0$）に分布していることが確認できる．Figure 3 のヒストグラムはコーパス別に色分けした曲率分布を示しており，ドメインによって曲率の分散に偏りがあるものの，いずれも 0 を閾値として左側（負曲率）に集中している．

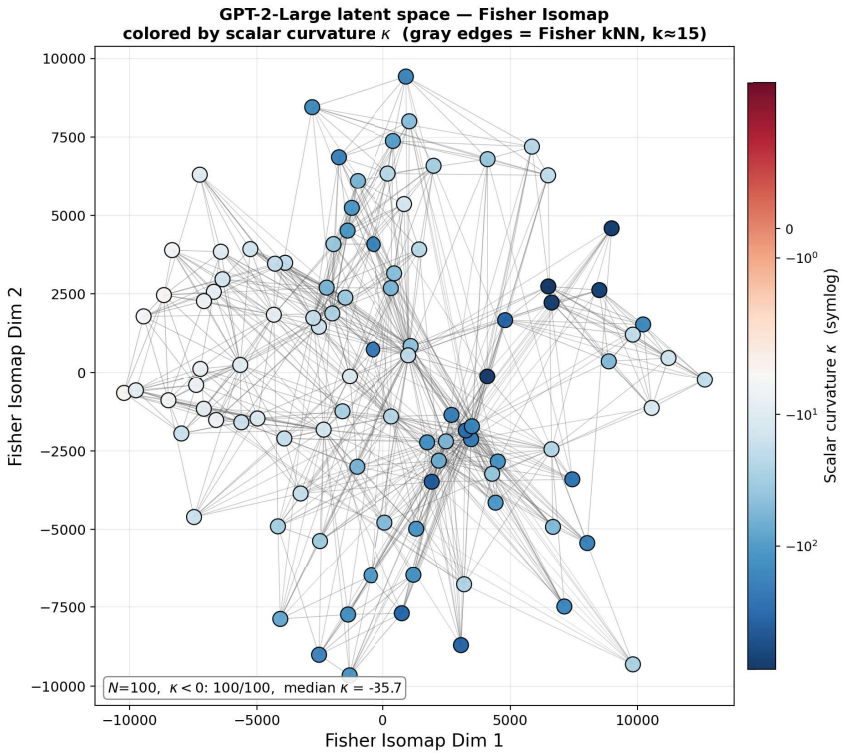

Figure 2: GPT-2-Large 潜在空間の Fisher Isomap 可視化．配置は 1280 次元 Fisher 計量 G_θ で重み付けた kNN グラフ ($k \approx 15$) 上の測地距離行列の MDS 埋め込み，色は PCA 上位 2 主成分が張る部分多様体上で評価したスカラー曲率 κ (負＝青，正＝赤，symlog スケール)．灰色のエッジは Fisher kNN.

幾何学的に $\kappa < 0$ とは双曲的な空間構造を意味する．双曲空間は階層的・ツリー状の構造を指数的な容量で効率的に埋め込む性質を持つため，この結果は言語モデルが知識を単なる点の集合ではなく，階層的に整理された概念構造として保持していることを示唆していると考えられる．

4.2 Fisher 測地距離による意味的類似度の実現

潜在空間において，単純なユークリッド距離 d_{euc} ではなく Fisher 計量 g_{ij} に基づく測地距離 d_{geo} を用いることで，人間が知覚する意味的な類似度が正しく定義されるかを検証した．画像認識モデル（MNIST-CNN）を用い，数字クラス「3」から「7」への最短経路（モーフィング）を，Vanilla Isomap（エッジ重みをユークリッド距離で固定）と Fisher Isomap（エッジ重みを Fisher 計量で重み付け）で比較した．

Vanilla Isomap では，最短経路が中間で混同しやすい他の数字クラス「1,2,5」のクラスタを横切る現象が観測された (Fig. 4 参照)．これは平坦な空間を仮定した直線距離が，多様体の局所的な歪みを無視した結果，意味的に不自然なクラス間ジャンプを誘発するためと考えられる．一方，Fisher Isomap では，計量 g_{ij} がデータ分散の逆行列として空間の縮尺を補正するため，経路が同一クラス内の多様体に沿って滑らかに遷移した．この結果は，測地距離こそが潜在空間における真の意味的近接性を捉える自然な尺度であることを示唆している．

4.3 LLM 間における概念空間の構造比較

次に，個別モデルの解析から複数モデル間の比較へと視点を広げる．異なるアーキテクチャ・スケールを持つ言語モデル群 (Mistral, LLaMA-2/3, CodeLLaMA, TinyLLaMA, Open-LLaMA, GPT-2

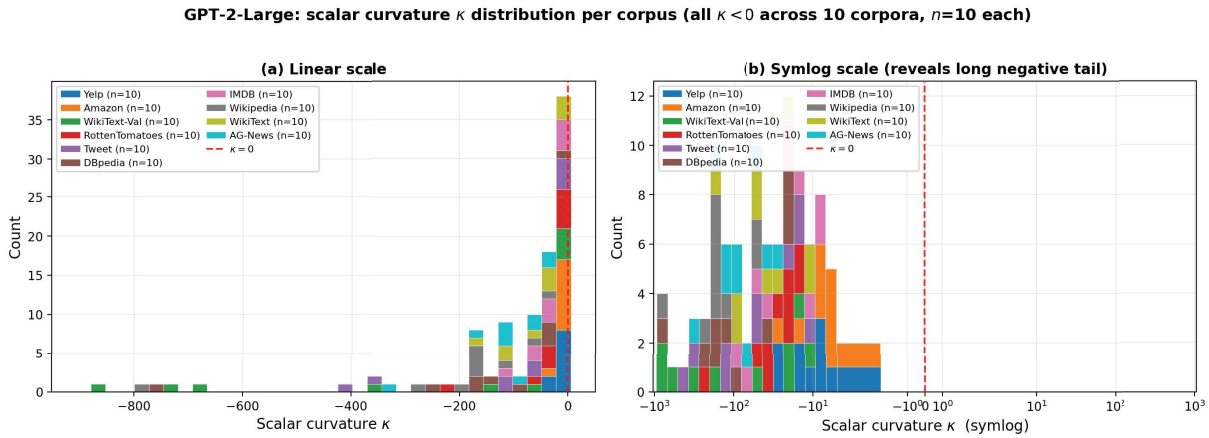

Figure 3: コーパス別スカラー曲率 κ の分布．全コーパスで負の領域に集中していることが確認される．

など 7 アーキテクチャ族・計 44 モデル) に共通の文書集合を入力し，各モデル M で得られる測地距離行列 $D^{(M)} \in \mathbb{R}^{N\times N}$ を算出した．モデル間の概念的乖離度を定量化するため，距離行列要素間の Spearman 順位相関 ρ を用いてモデル間距離を定義する：

$$d_{\text{model}}(M_a, M_b) = 1 - \rho\big(D^{(M_a)}, D^{(M_b)}\big).$$

この距離行列に対して多次元尺度構成法（MDS）を適用し，2 次元平面に投影した「LLM の地図」(Fig. 5) を作成した．

解析結果は以下の知見を示した．第一に，同一アーキテクチャ族に属するモデルは地図上で密にクラスター化し，世代間の差 (例：LLaMA-2 ↔ LLaMA-3) は明確な距離分離として現れた．第二に，モデルの位置を決定づけている要因は，パラメタ数や隠れ層次元といったスペック的な指標よりも，アーキテクチャの設計原理や学習データの系譜であることが判明した (調整 Rand 指数 ARI = 0.659)．これは，外部の言語性能スコアではなく，内部中間表現がなす幾何構造そのものがモデルの「概念的枠組み」を特徴付けると捉えることができる．本手法は，基盤モデルのブラックボックス性を幾何学的指標で透過化する有効な手段であることを示している．

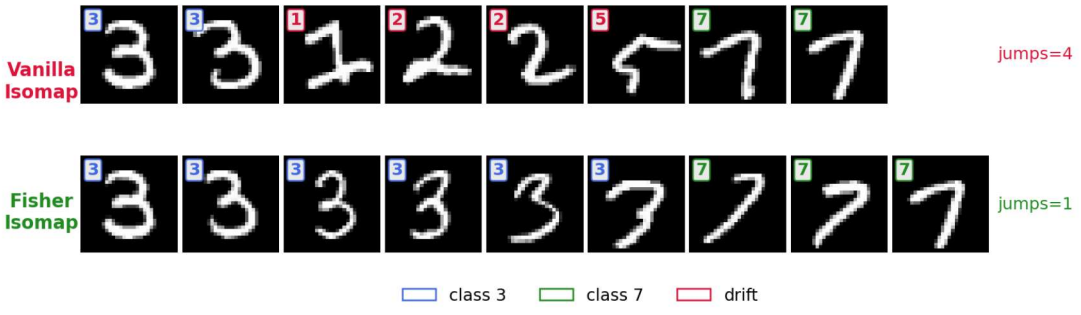

Figure 4: 数字クラス「3」から「7」へのモーフィング経路の比較．上段: Vanilla Isomap (ユークリッド距離) — 中間で他クラス「1,2,5」を経由 (class jumps = 4)．下段: Fisher Isomap (Fisher 計量) — クラス間で多様体に沿って滑らかに遷移 (class jumps = 1)．色つきバッジは各画像のクラスラベル (青=source「3」，緑=target「7」，赤=他クラス drift)．

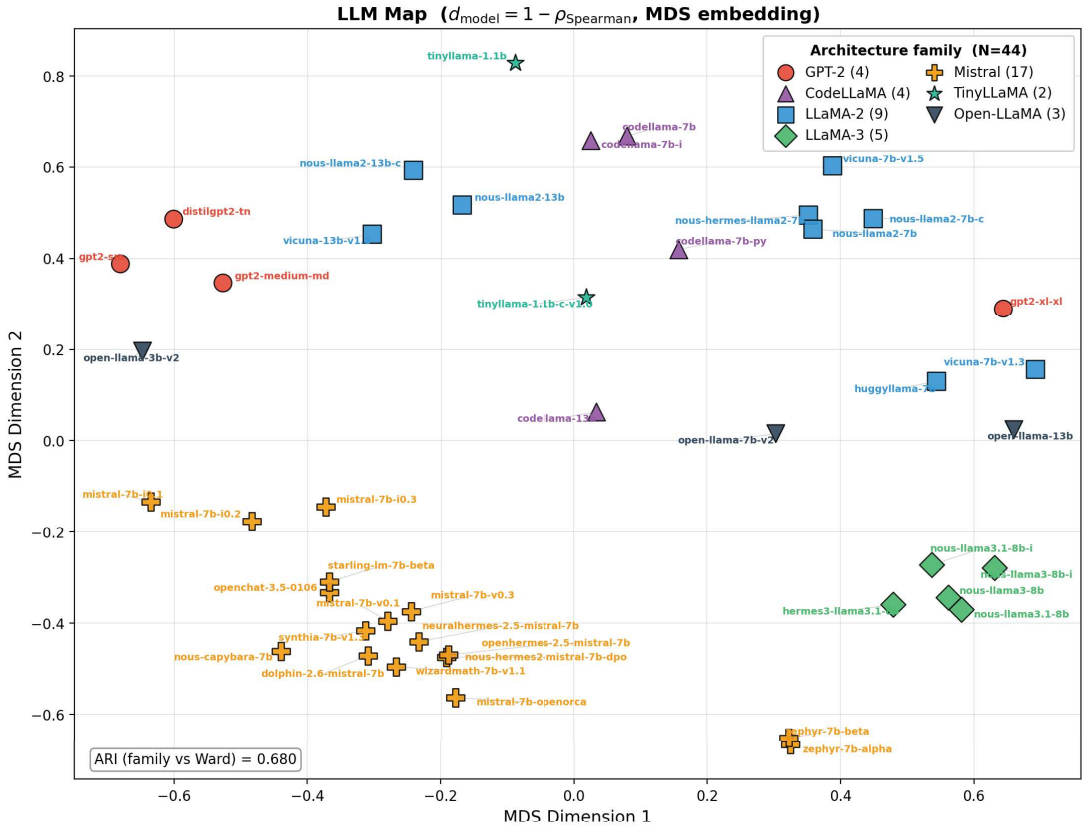

Figure 5: Spearman 順位相関距離に基づく LLM 地図 (44 モデル，7 アーキテクチャ族の MDS 埋め込み)．色とマーカー形状はアーキテクチャ族 (GPT-2, CodeLLaMA, LLaMA-2, LLaMA-3, Mistral, TinyLLaMA, Open-LLaMA) を示す．ARI (族 vs Ward 階層クラスタリング) = 0.659．

5 おわりに

本稿では，基盤モデルの出力 $\eta(x)$ を指数型分布族の期待値パラメタとして解釈し，その分布族 $\mathcal{M}$ の幾何構造を情報幾何の枠組みで考察した．具体的には，モデル内部から抽出可能な潜在表現と重みを標本とみなすことで，Fisher 計量 g_{ij}，アファイン接続 Γ_{ijk}，およびスカラー曲率 κ を標本統計量から推定した．この枠組みを一部の基盤モデルにおいて実証した結果，LLM では潜在空間が全体的に負の曲率を示す双曲的な性質を有することがわかり，モデルが知識を階層的な構造として保持している可能性が示唆された．また，Fisher 計量に基づく測地距離により，ユークリッド距離で生じる不自然なクラス間遷移を抑制し，意味的にも滑らかなモーフィングを実現できた．さらに，複数 LLM 間の測地距離行列の相関解析から作成した LLM 地図では，モデルの幾何構造がパラメタ規模よりも学習器の構造や学習データの系譜を強く反映していることが確認された．

従来の学習モデルの評価は，パラメタの値や外部出力，あるいは特定のタスクにおける正解率といった結果としての挙動や静的な数値に依存することが多かったが，最近になり深層学習モデルの内部構造を幾何的・統計的に理解する新たな試みはいくつか提案されている．例えば Khemakhem ら [12] は，本稿と同様に学習機械を指数型分布族と対応づけて，潜在空間の性質を議論している．こうした試みは，モデルが内部的に構築している「概念空間の構造」という，より抽象度の高い次元でモデルを定量化し，モデルの規模や具体的な出力値に依存することなく，そのモデルが世界をどのような形で捉えているかという，いわば「モデルの個性」や「認知的な枠組み」を客観的に記述しようとするものと言えよう．

今後の課題としては，まず学習ダイナミクスの幾何学的解析が挙げられる．モデルの学習過程に

おいて，潜在空間の計量や曲率がどのように時間発展するかを調べることにより，どのようなプロセスを経て概念的な階層構造が形成されていくかを解き明かすことができるかもしれない．また，同一のアーキテクチャを用いながら学習データの内容や質を変えた場合に，空間の形状がどう変化するかを調査することで，データの性質と幾何構造の関係を明らかにすることも重要な課題である．さらに，潜在空間のスカラー曲率などの幾何学的指標が，モデルの実際の推論能力や汎化性能にどのような影響を与えるかを検証することも必要であろう．ブラックボックス化しがちな基盤モデルの内部構造に対し，情報幾何の視点から潜在空間の「形」を理解することが，より効率的で解釈可能なモデルを設計するための一つの指針となることを期待したい．

参考文献

1. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE* **86,** 2278–2324 (Jan. 1, 1998).
2. Vaswani, A. *et al. Attention Is All You Need* in *Advances in Neural Information Processing Systems* **30** (Curran Associates, Inc., 2017).
3. Dosovitskiy, A. *et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale* in. International Conference on Learning Representations (Oct. 2, 2020).
4. Tenenbaum, J. B., Silva, V. de & Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **290,** 2319–2323 (Dec. 22, 2000).
5. Amari, S.-i. *Differential-Geometrical Methods in Statistics* (red. Berger, J. *et al.*) (Springer, New York, NY, 1985).
6. 赤塚, 育. & 村田, 昇. エンコーダの潜在空間の幾何学量推定. **人工知能学会全国大会論文集 JSAI2025,** 3S1GS204–3S1GS204 (2025).
7. Radford, A. *et al.* Language Models Are Unsupervised Multitask Learners. *OpenAI blog* **1,** 9 (2019).
8. Lhoest, Q. *et al. Datasets: A Community Library for Natural Language Processing* in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (eds Adel, H. & Shi, S.) (Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, Jan. 1, 2021), 175–184.
9. Merity, S., Xiong, C., Bradbury, J. & Socher, R. *Pointer Sentinel Mixture Models* in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* (OpenReview.net, 2017).
10. Maas, A. L. *et al. Learning Word Vectors for Sentiment Analysis* in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* (Association for Computational Linguistics, USA, June 19, 2011), 142–150.
11. Zhang, X., Zhao, J. & LeCun, Y. *Character-Level Convolutional Networks for Text Classification* in *Advances in Neural Information Processing Systems* **28** (Curran Associates, Inc., 2015).
12. Khemakhem, I., Kingma, D., Monti, R. & Hyvarinen, A. *Variational Autoencoders and Nonlinear ICA: A Unifying Framework* in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* International Conference on Artificial Intelligence and Statistics (PMLR, June 3, 2020), 2207–2217.

Faculty of Sciences and Engineering
Waseda University
Tokyo 169-8555, JAPAN
E-mail address: noboru.murata@eb.waseda.ac.jp

早稲田大学・理工学術院 村田 昇