系統樹にまつわる幾何数理工学

平井広志
東京大学工学部 計数工学科 数理情報工学コース
東京大学大学院 情報理工学系研究科 数理情報学専攻
hirai@mist.i.u-tokyo.ac.jp
協力：池田基樹（数理情報学専攻 D1）

0 準備：距離空間

(X,d) が**距離空間**とは，$d : X \times X \to \mathbb{R}$ が**距離関数（メトリック）**と呼ばれる次の性質を満たす関数であることをいう．

- $d(x,y) = 0 \Leftrightarrow x = y$.
- $d(x,y) = d(y,x)$.
- $d(x,y) + d(y,z) \geq d(x,z)$.

定義 1. 連続写像 $P : [0,1] \to X$ のことを**パス**という．P は $x = P(0)$ と $y = P(1)$ を結ぶと言ったりする．パス P の長さ $d(P)$ は

$$d(P) = \sup \sum_{i=1}^{N} d(P(t_{i-1}), P(t_i)) \tag{1}$$

と定義される．ここで sup は $N > 0$ と $0 = t_0 < t_1 < \cdots < t_N = 1$ にわたってとる．

常に $d(P) \geq d(x,y)$ が成り立つことに注意する．

定義 2. 距離空間 X が**測地的** *(geodesic)* とは，任意の $x,y \in X$ に対して x と y を結ぶパス P で，$d(x,y) = d(P)$ を満たすものが存在することをいう．このようなパス P で，弧長に比例するパラメトライゼーションをとったものを**測地線**という．つまり，$d(P(s),P(t)) = |s-t|d(x,y)$ $(s,t \in [0,1])$.

例 1. ユークリッド空間 $\mathbb{R}^n$ は測地的.

例 2 (ツリー，木)**.** 有限の線分 ($\simeq [0,a]$ $(a > 0)$) を，サイクルができないように（連結になるように）端点で貼り合わせたものを**ツリー**，**木**という（図 1）．パス P の長さ $d(P)$ を (1) で定義する．ここで，$P(t_{i-1}), P(t_i)$ は同じ線分に含まれるようにとり，$d(P(t_{i-1}), P(t_i))$ はその線分内で測る．距離関数 d を

$$d(x,y) = \inf_{P:(x,\,y)\text{-パス}} d(P)$$

と定義すると，(X,d) は測地的距離空間.

例 3 (立方複体)**.** いくつかのキューブ $[0,1]^k \subseteq \mathbb{R}^k$ を面で貼り合わせたものを**立方複体**という（図 2）．ツリーのときと同様にパスの長さと距離関数 d を定めると，(X,d) は測地的距離空間.

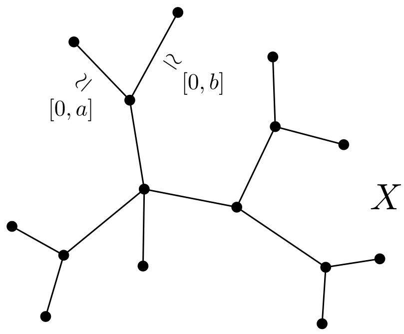

図 1: ツリーの例.

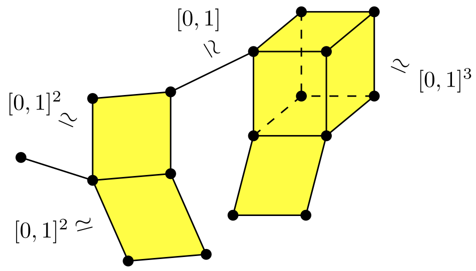

図 2: 立方複体の例.

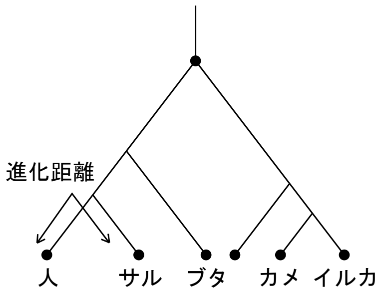

図 3: 系統樹.

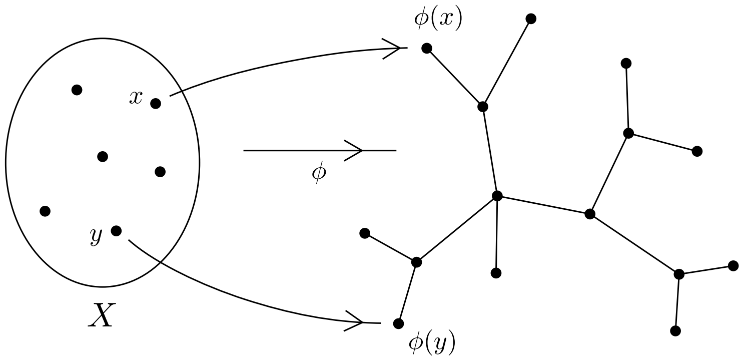

図 4: 木距離.

1　系統樹，木距離，4 点条件

与えられた生物種の DNA データから系統樹（図 3）を復元したい．距離法と呼ばれる手法では，DNA データから各生物種間の距離 d を求め，d に「フィットする」系統樹をつくる．そのためには，ツリーに埋め込める距離を特徴づける必要がある．そのような距離を**木距離** *(tree metric)* という．

定義 3. (X,d) を有限距離空間とする（つまり X が有限集合）．d が**木距離**とは，（例 2 の意味での）ツリー T が存在し，$\phi : X \to T$ で

$$d(x,y) = d_T(\phi(x),\phi(y)) \quad (x,y \in X)$$

を満たすものが存在することをいう（図 4）．

定理 4 (Buneman, 71)**.** 以下は同値.

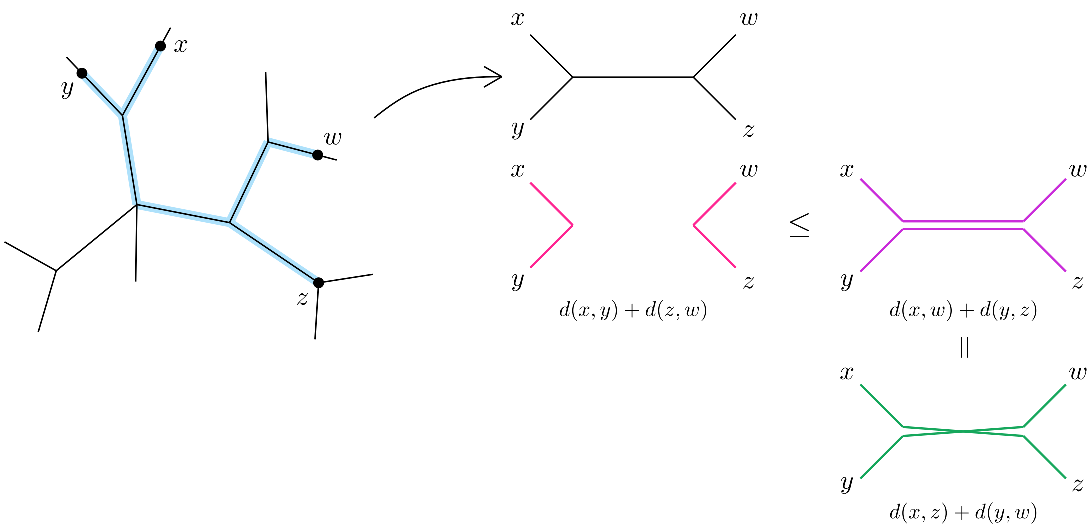

図 5: 4 点条件.

(i) d が木距離.

(ii) d は以下の条件を満たす（4 点条件（four point condition）と呼ばれる）.

$$d(x,y)+d(z,w) \leq \max\{d(x,z)+d(y,w), d(x,w)+d(y,z)\} \quad (\forall x,y,z,w \in X).$$

つまり，$d(x,y)+d(z,w)$，$d(x,z)+d(y,w)$，$d(x,w)+d(y,z)$ のうち，最大を達成するものが 2 個以上ある．

(iii) ラミナー族 $\mathcal{L} \subseteq 2^X$ と $\alpha : \mathcal{L} \to \mathbb{R}_{++}$ が存在し，

$$d = \sum_{S \in \mathcal{L}} \alpha(S)\delta_S$$

が成り立つ．ここで δ_S は

$$\delta_S(x,y) = \begin{cases} 0 & \text{if } x,y \in S \text{ or } x,y \notin S, \\ 1 & \text{o.w. } (x \in S \not\ni y \text{ or } x \notin S \ni y) \end{cases}$$

で定義され，カットセミメトリックと呼ばれる．$\mathcal{L} \subseteq 2^X$ がラミナーとは，$\forall S,T \in \mathcal{L}$ が $S \cap T = \emptyset$，$S \subseteq T$，$S \supseteq T$ のどれかを満たすことをいう．

証明の気分. (i) $\Rightarrow$ (ii) ツリーの任意の 4 点に対してそれらを含む最小のツリーを取ると，一般性を失わず図 5 の右上のような距離を仮定できる．すると，

$$d(x,y)+d(z,w) \leq d(x,w)+d(y,z) = d(x,z)+d(y,w)$$

となる．

(i) $\Leftrightarrow$ (iii) ツリーの根を任意に固定する．各枝 e に対して，e を取り除いて出来る 2 つの連結成分のうち根を含まないほうに（イメージが）含まれる集合 $S_e \subseteq X$ を対応させる（図 6）．$\alpha(S_e)$ を e の長さとする．すると $d = \sum_{e:枝} \alpha(S_e)\delta_{S_e}$ で，$\{S_e \mid e : 枝\} \subseteq 2^X$ はラミナー．

□

2 タイトスパン (Tight span, Dress, 84)

タイトスパンとは，任意のメトリック d を表現する「幹のある」系統樹のことである．d が木距離なら，それを表現するツリーになる．

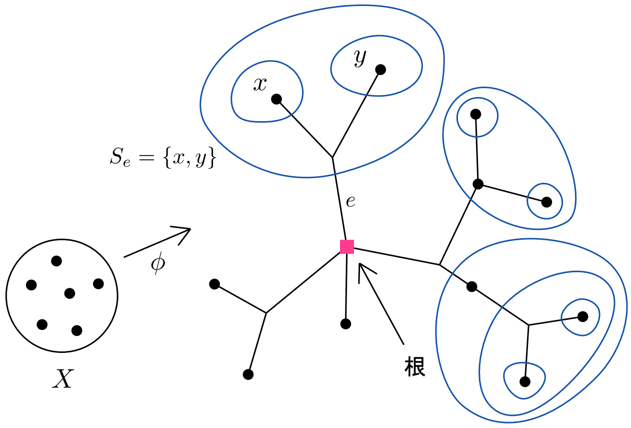

図 6: ラミナー族の構成.

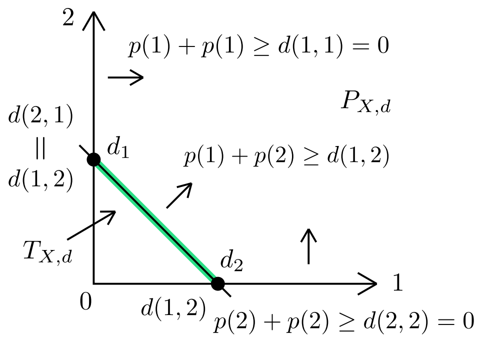

図 7: $X = \{1, 2\}$ の場合のタイトスパン.

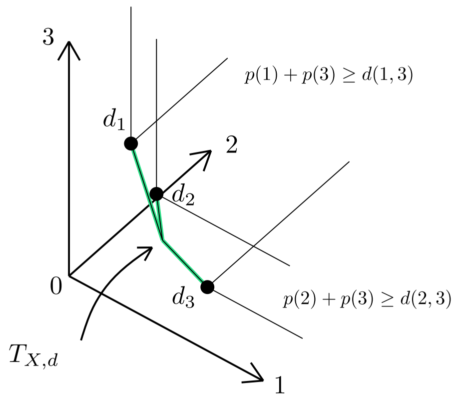

図 8: $X = \{1, 2, 3\}$ の場合のタイトスパン.

(X, d) を有限距離空間とする．多面体 $P_{X,d}$ を

$$P_{X,d} := \{p \in \mathbb{R}_+^X \mid p(x) + p(y) \geq d(x, y),\ x, y \in X\}$$

と定義する．

定義 5 (タイトスパン)**.** タイトスパン $T_{X,d}$ を，$P_{X,d}$ の極小な点全体と定義する．ここで $p \in P_{X,d}$ が極小とは，$p' \in P_{X,d}$, $p' \leq p$ $(p'(x) \leq p(x)$ $(\forall x \in X))$ なら $p' = p$ であることをいう．定義より $p \in P_{X,d}$ が極小 $\Leftrightarrow \forall x \in X,\ \exists y \in X,\ p(x) + p(y) = d(x, y)$.

例 4. $X = \{1, 2\}$ の場合のタイトスパンの例を図 7 に示す．$P_{X,d}$ は直線 $p(1) = 0$, $p(2) = 0$, $p(1)+p(2) = d(1, 2)(= d(2, 1))$ で囲まれた領域になる．よってタイトスパン $T_{X,d}$ は線分 $\{p \in \mathbb{R}_+^X \mid p(1) + p(2) = d(1, 2)\}$ になる．

$X = \{1, 2, 3\}$ の場合のタイトスパンの例を図 8 に示す．$P_{X,d}$ は平面 $p(1) = 0$, $p(2) = 0$, $p(3) = 0$, $p(1) + p(2) = d(1, 2)$, $p(2) + p(3) = d(2, 3)$, $p(1) + p(3) = d(1, 3)$ で囲まれた領域になる．よってタイトスパン $T_{X,d}$ は図 8 で緑に塗られた領域になる．

$X = \{1, 2, 3, 4\}$ の場合のタイトスパン $T_{X,d}$ は図 9 のような領域になる．

ℓ_∞-距離でタイトスパンを距離空間にすることができる．つまり，$p, q \in T_{X,d}$ に対し，

$$d_{T_{X,d}}(p, q) := \|p - q\|_\infty = \max_{x \in X} |p(x) - q(x)|$$

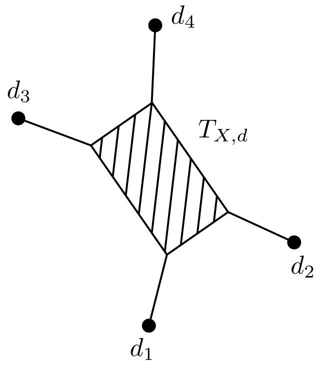

図 9: $X = \{1, 2, 3, 4\}$ の場合のタイトスパン.

と定義する．X の各点をタイトスパン $T_{X,d}$ に等長埋め込みできる．$x \in X$ に対し，$d_x \in \mathbb{R}^X$ を

$$d_x(y) := d(x, y) \quad (y \in X)$$

とおく．例えば $X = \{1, 2, \dots, n\}$ なら

$$d_x = (d(x, 1), d(x, 2), \dots, d(x, n))^\top.$$

補題 6. (i) $d_x \in T_{X,d}$ $(x \in X)$.

(ii) $\|d_x - d_y\|_\infty = d(x, y)$ $(x, y \in X)$.

つまり，(X, d) は $x \mapsto d_x$ によって $(T_{X,d}, \ell_\infty)$ に等長的に埋め込まれる．

証明. (i) $d_x(y) + d_x(z) = d(x, y) + d(x, z) \geq d(y, z)$ より $d_x \in P_{X,d}$ となる．また任意の $y \in X$ に対し，$d_x(x) + d_x(y) = 0 + d_x(y) = d(x, y)$ より d_x は極小.

(ii) $|d_x(x) - d_y(x)| = |0 - d(y, x)| = d(y, x)$ と $|d_x(z) - d_y(z)| = |d(x, z) - d(y, z)| \leq d(x, y)$ より $\|d_x - d_y\|_\infty = d(x, y)$ を得る． □

定理 7 (Dress, 84)**.** (i) $T_{X,d}$ は可縮.

(ii) $T_{X,d}$ は測地的.

(iii) d が木距離 $\Leftrightarrow \dim T_{X,d} = 1$ （厳密には ≤ 1).

ここで $\dim T_{X,d}$ は，$T_{X,d}$ を構成する多面体の最大次元を表す．つまり $\dim T_{X,d} = 1$ は $T_{X,d}$ がツリーであることを意味する.

補題 8 (Dress's retraction lemma)**.** $\phi : P_{X,d} \to T_{X,d}$ が存在し，以下の条件を満たす.

(i) $\phi(p) = p$ $(p \in T_{X,d})$.

(ii) $\|\phi(p) - \phi(q)\|_\infty \leq \|p - q\|_\infty$ $(p, q \in P_{X,d})$.

証明. $\phi_x : P_{X,d} \to P_{X,d}$ を，$p \in P_{X,d}$ に対し，p の x 成分をできるかぎり減少させた要素を対応させる写像と定義する $(p(x) \mapsto p(x) - \alpha)$．すると ϕ_x は (ii) を満たす．実際, 一般性を失わず $\phi_x(p)(x) - \phi_x(q)(x) \geq 0$ と仮定したとき，$y (\neq x)$ が存在して

$$\begin{aligned}
\phi_x(p)(x) - \phi_x(q)(x) &= d(x, y) - p(y) - \phi_x(q)(x) \\
&\leq d(x, y) - p(y) - (d(x, y) - q(y)) \\
&\leq -p(y) + q(y) \leq \|p - q\|_\infty
\end{aligned}$$

となる．この写像 ϕ_x を X の全要素分だけ合成してできる写像を考える．すなわち，$X = \{1, 2, \dots, n\}$ なら

$$\phi = \phi_n \circ \phi_{n-1} \circ \cdots \circ \phi_1$$

とすれば，所望の $\phi : P_{X,d} \to T_{X,d}$ を得る． □

定理 7 (i) の証明. $(t,x) \mapsto t\phi(x) + (1-t)x$ と変形レトラクションが構成できる．よって $T_{X,d}$ は $P_{X,d}$ にホモトピー同値．$P_{X,d}$ は凸であるから可縮． □

定理 7 (ii) の証明. $p, q \in T_{X,d}$ とする．P を任意の $T_{X,d}$ 内の (p,q)-パスとすると，$d(P) \geq \|p-q\|_\infty$ になることは明らか．P を $P_{X,d}$ 内の (p,q)-測地線とすると P は直線で，$d(P) = \|p-q\|_\infty$ となる．これを ϕ によって写した $\phi(P)$ は，$T_{X,d}$ 内での (p,q)-パスになる．補題 8 より $d(\phi(P)) \leq d(P) = \|p-q\|_\infty$ なので，$d(\phi(P)) = \|p-q\|_\infty$． □

定理 7 (iii) の証明. $\dim T_{X,d} = 1$ なら (i) より $T_{X,d}$ はツリー．補題 6 と (ii) より d は木距離．

d を木距離とすると 4 点条件を満たす．$p \in T_{X,d}$ を任意に取り，

$$E(p) := \{(x,y) \in X \times X \mid p(x) + p(y) = d(x,y)\}$$

とおく．行列 $M \in \mathbb{R}^{X \times (X \times X)}$ を

$$M_{xe} := \begin{cases} 1 & \text{if } e = xy \in E(P), \\ 0 & \text{o.w.} \end{cases}$$

と定義する．$\dim\{q \in \mathbb{R}^X \mid q^\top M = 0\} \leq 1$ を示せばよい．

$$\operatorname{rank} M^\top = \operatorname{rank} M = |X| - (X, E(P)) \text{ の 2 部連結成分数}$$

なので，$\dim\ker M^\top$ は $(X, E(P))$ の 2 部連結成分数に一致する．もしもこれが ≥ 2 なら，ある $xy, zw \in E(P)$ があって異なる成分に属している．しかし

$$\begin{aligned} &p(x) + p(y) = d(x,y), \quad p(x) + p(z) > d(x,z), \quad p(x) + p(w) > d(x,w), \\ &p(z) + p(w) = d(z,w), \quad p(y) + p(w) > d(y,w), \quad p(y) + p(z) > d(y,z) \end{aligned}$$

は $d(x,y) + d(z,w) = p(x) + p(y) + p(z) + p(w) > \max\{d(x,z) + d(y,w), d(x,w) + d(y,z)\}$ を導き，4 点条件に矛盾する． □

3　系統樹空間（Billera, Holmes, Vogtmann, 2001）

系統樹（∼ 木距離）全体を距離空間にする．

3.1　準備：CAT(0) 空間（Gromov, 1987）

CAT は Cartan, Alexandrov, Topogonov の頭文字，0 は曲率が 0 以下であることを意味する．

X を測地的距離空間とする．$x, y, z \in X$ をとると，各 2 点の間に測地線が引ける．これを測地的三角形（図 10 左）という．これに対し，$\mathbb{R}^2$ の上に

$$\begin{aligned} d(x,y) &= \|\bar{x} - \bar{y}\|_2, \\ d(y,z) &= \|\bar{y} - \bar{z}\|_2, \\ d(z,x) &= \|\bar{z} - \bar{x}\|_2 \end{aligned}$$

を満たす比較三角形 $\bar{x}\bar{y}\bar{z}$ が取れる（図 10 右）．このような三角形は合同，反射を除いてユニークに存在する．すると，測地的三角形の辺上の点 p に対し，対応する比較三角形上の点 $\bar{p}$ が決まる．

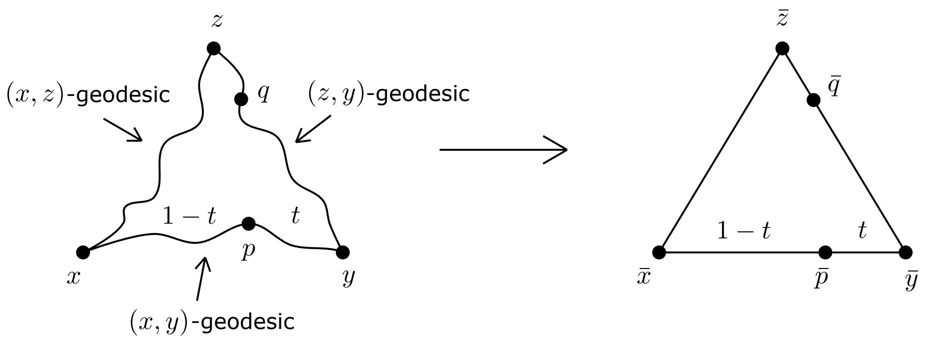

図 10: 測地的三角形と比較三角形.

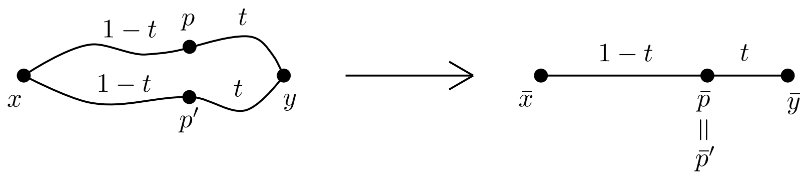

図 11: CAT(0) 空間は一意測地的.

定義 9 (Gromov)**.** x が CAT(0) $\overset{\text{def}}{\Longleftrightarrow}$ 任意の測地的三角形とその任意の辺上の 2 点 p, q に対し，$d(p,q) \leq \|\bar{p} - \bar{q}\|_2$．つまり，三角形が痩せている．

例 5. $\mathbb{R}^n$ やツリー，双曲空間は CAT(0) 空間.

命題 10. X を CAT(0) 空間とする．

(i) X は一意測地的（$\overset{\text{def}}{\Longleftrightarrow}$ 任意の 2 点を結ぶ測地線がユニークに決まる）．

(ii) X は可縮.

証明. (i) x, y の間に測地線が 2 本あるとする．これを測地的三角形 x, x, y と見ると，対応する比較三角形は $\bar{x}$ と $\bar{y}$ を結ぶ直線になる（図 11）．よって，x, y の間の 2 本の測地線を $1-t : t$ の比率で分ける点をそれぞれ p, p' とすると，$\bar{p} = \bar{p}'$ となる．CAT(0) 性より $d(p,p') \leq \|\bar{p} - \bar{p}'\| = 0$．よって $p = p'$．

(ii) $z \in X$ を任意にとる．$(x,t) \mapsto p_x(t)$（p_x は (x,z)-測地線）は一点 z への変形レトラクションになる．連続性は次のように確かめられる（図 12）．$\mathbb{R}^2$ で $\bar{x} \mapsto (1-t)\bar{x} + t\bar{z}$ は連続であるから，任意の $\epsilon > 0$ に対して $\delta > 0$ が存在し，$(\bar{x}, t)$ の δ-近傍内の点 $(\bar{y}, s)$ が $\|((1-t)\bar{x} + t\bar{z}) - ((1-s)\bar{y} + s\bar{z})\|_2 < \epsilon$ を満たすようにできる．すると，$X \times \mathbb{R}$ において (x,t) の δ-近傍内の点 (y,s) を取れば，対応する点 $(\bar{y}, s)$ は $(\bar{x}, t)$ の δ-近傍となる．測地的三角形 x, y, z を考えれば

$$d(p_x(t), p_y(s)) \leq \|((1-t)\bar{x} + t\bar{z}) - ((1-s)\bar{y} + s\bar{z})\|_2 < \epsilon.$$

□

3.2 CAT(0) 立方複体の組合せ的特徴付け

定理 11 (Gromov, 1987)**.** X を立方複体とする．以下は同値.

- X は CAT(0).
- X は単連結で各頂点のリンクが**フラッグ**.

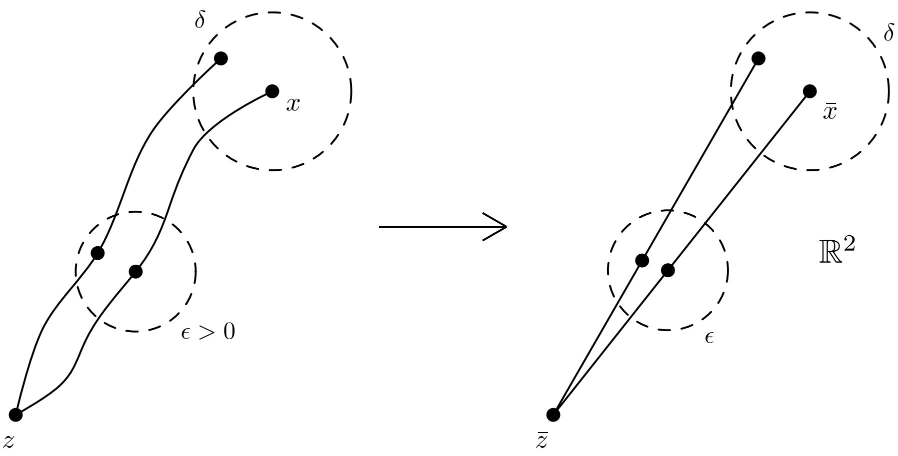

図 12: CAT(0) 空間は可縮.

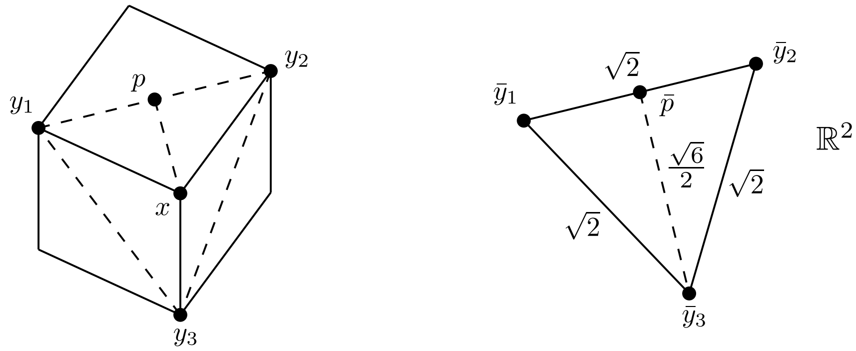

図 13: リンクがフラッグでない立方複体の例.

ここで頂点 x のリンクが**フラッグ**とは，x と隣接する頂点 $y_1, y_2, \ldots, y_k$ を含む cube が存在することを言う．これは，任意の ij について x と y_i，y_j を含む cube が存在することと同値である．

フラッグでない例を図 13 に示す．頂点 x を図のように選ぶと，隣接する頂点 y_1, y_2, y_3 を全て同時に含む cube は存在しない．この立方複体が CAT(0) でないことも容易に確かめられる．測地的三角形 y_1, y_2, y_3 と，対応する $\mathbb{R}^2$ 上の比較三角形 $\bar{y}_1, \bar{y}_2, \bar{y}_3$ を取る．y_1 と y_2 の中点を p，比較三角形上の対応する点を $\bar{p}$ とおくと，$d(p, y_3) = \sqrt{2}/2 + 1 = 1.7\cdots$ なのに対し，$\|\bar{p} - \bar{y}_3\|_2 = \sqrt{6}/2 = 1.2\cdots$ で，$d(p, y_3) > \|\bar{p} - \bar{y}_3\|_2$ となる．