単語ベクトルの平行四辺形を特徴づける図形距離

前田晃弘 [1]　鳥居拓馬 [2]　日髙昇平 [1]
[1] 北陸先端科学技術大学院大学　[2] 東京電機大学
{akihiro.maeda,shhidaka}@jaist.ac.jp　tak.torii@mail.dendai.ac.jp

概要

単語の意味的・統語的関係を表わす単語ベクトルが四項類推課題を解くことはよく知られている．そのベクトル演算は類推関係にある単語ベクトルが平行四辺形をなしていることを示唆している．これに着目して高次元空間にある単語ベクトルの配置を図形の幾何的性質により特徴づけることを試みる．四つの単語からなる単語群の平行四辺形らしさを測定するメトリックとして図形距離を提案し，これを用いて類推関係にある単語群が平行四辺形に近い配置にあることを実証する．図形距離は単語群の関係性を定量的かつ精緻に特徴づける性能を有しており，新たな研究ツールとなるとともに言語構造の解明に幾何的にアプローチすることを可能にする．

1　はじめに

1.1　単語ベクトルによる四項類推課題

単語埋め込み（単語ベクトル）は単語間の意味的・統語的関係を表していることが知られ，四項類推課題はその代表的な現象である．例えば，*king:queen::man:woman* のような類推関係に対して，それぞれの単語ベクトル間に $\boldsymbol{v}_{king}-\boldsymbol{v}_{queen}+\boldsymbol{v}_{woman} \approx \boldsymbol{v}_{man}$ という演算が成り立つ．Mikolov et al. [1] 等の先行研究では典型的にコサイン類似度により両辺が比較されるが，同時に *king:man::queen:woman* の語順の平行関係も成り立つことから四項類推関係は平行四辺形の存在を示唆すると考えられる [2, 3].

1.2　平行四辺形は存在するか？

単語ベクトルは高次元空間の点であるが，その 4 点が平行四辺形のような平面図形をなしているとはどの様な状態であろうか．ユークリッド距離やコサイン類似度は 2 つの単語ベクトル間の関係性を評価するものであるが，3 以上の単語間にも関係性や対称性があると考えられる．また構造言語学の立場からは単語の意味関係は絶対的なものではなく差異により特徴づけられるとされる [4]．図形のもつ幾何的性質を用いて高次元空間に住む単語ベクトルの集まりが持つ関係性を特徴づけたいというのが本研究の動機である．

1.3　単語関係を幾何的に捉える

本研究では類推関係にある単語ベクトルが示唆する平行四辺形に着目して，まず高次元ベクトル空間における平行四辺形を数学的に定義する．その上で四つの単語ベクトルが平行四辺形をなす条件を導出して，平行四辺形らしさを表す新しい指標として図形距離を提案する．さらに図形距離を実コーパスから生成された単語ベクトルに適用してその幾何的性質を定量的に評価する．単語ベクトル間の関係を幾何的に捉えることで言語の豊かな対称性を抽出する．

2　図形距離

2.1　高次元空間の平行四辺形

ユークリッド幾何における図形とは任意の相似変換によって不変に保たれる幾何学的性質を持つ対象として定義される [5]．従って，高次元空間内の図形も相似変換群の全ての変換に対して不変に保たれる性質によって定義できる．

平面における平行四辺形は複数の性質[1)]を持つが，本論文では高次元ベクトル空間における平行四辺形を次のように定義する．

定義　d 次元ベクトル空間における 4 つのベクトル $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ について，

$$\mathbf{x}_1 - \mathbf{x}_2 = \mathbf{x}_3 - \mathbf{x}_4 \tag{1}$$

が成り立つ時 4 つのベクトルは平行四辺形をなす．

1)　平行四辺形は (i)2 組の対辺がそれぞれ平行，(ii)1 組の対辺が平行かつ同じ長さ，(iii)2 組の対角がそれぞれ等しい，(iv) 対角線の中点が一致するという性質を持つ．

2.2 頂点の選び方

任意に与えられた 4 つのベクトルを定義の式 (1) にある $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ へ割り当てる選び方は $4! = 24$ 通りあり，頂点の置換は 4 次対称群 S_4 をなす．このうち式 (1) を不変とする置換は，ベクトルの添字を置換の対象として表せば $\{(14), (23), (12)(34)\}$ を生成元とする S_4 の部分群である．位数は $2 \times 2 \times 2 = 8$ であるので，4 つの頂点に対して異なる四角形をなす頂点の選び方は $24/8 = 3$ 通り存在する．

図 1 は，3 通りの頂点の選び方に対応した平行四辺形を表しており，パタン 1 は $\mathbf{x}_1$ と $\mathbf{x}_4$ が対角となる頂点の選び方であり，パタン 2 では $\mathbf{x}_1$ と $\mathbf{x}_3$ が対角，パタン 3 では $\mathbf{x}_1$ と $\mathbf{x}_2$ が対角をなす．

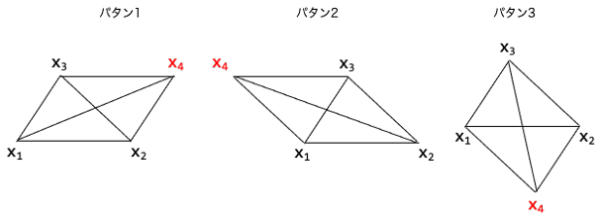

図 1　四角形の頂点の選び方

2.3 平行四辺形の必要十分条件

4 つの単語 w_1, w_2, w_3, w_4 を表す単語ベクトルをそれぞれ $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4 \in \mathbb{R}^d$ とし，行方向に沿って並べた行列 $X = [\mathbf{x}_1\ \mathbf{x}_2\ \mathbf{x}_3\ \mathbf{x}_4]$ を単語ベクトル群と呼ぶ．X は d 行 4 列の行列である．このとき単語ベクトル群 X が定義 (1) を満たすための必要十分条件は，3 つの頂点の選び方を考慮して

$$\mathbf{p}_1 = \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}, \quad \mathbf{p}_2 = \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}, \quad \mathbf{p}_3 = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix} \tag{2}$$

とすると，

$$X\mathbf{p}_1 = \mathbf{0}_d, \quad X\mathbf{p}_2 = \mathbf{0}_d, \quad X\mathbf{p}_3 = \mathbf{0}_d \tag{3}$$

のいずれかが充されることである．$\mathbf{0}_d$ は d 次元のゼロベクトルである[2]．

式 (3) の形より 4 つのベクトルは d 次元空間内の超平面上にあることが直ちにわかる．

2)　単語ベクトル間の関係性を行列により表現することの利点は，幾何学・線形代数・群論などの数学的操作を言語構造に対して行うことが可能となることである [6].

2.4 平行四辺形らしさの定量化

次に単語ベクトル群の平行四辺形までの距離を定める．考え方としては，単語ベクトル群 X を d 次元空間 V 上の 4 つのベクトルの集まりではなく，$4d$ 次元空間 W 上の 1 点 $\mathbf{x} := \mathrm{vec}(X) \in W$ とする[3]．平行四辺形をなしている単語ベクトル群を表すすべての点の集まりを W の部分空間 U とすれば，点 $\mathbf{x}$ から部分空間 U までの距離（ユークリッド距離）を算出でき，これを平行四辺形までの距離（図形距離）と定義できる．

図形距離がゼロであるとき単語ベクトル群は平行四辺形をなし，図形距離（正値）が大きくなるほど平行四辺形から遠くなる．平行四辺形をなす全ての単語ベクトル群の集合は，W の部分空間（カーネル）K_i として次式で表される[4]．

$$K_i = \{X \in V^4 | X\mathbf{p}_i = \mathbf{0}_d\} \tag{4}$$

$$= \{\mathbf{x} \in W | (\mathbf{p}_i^t \otimes I_d)\mathbf{x} = \mathbf{0}_d\} \quad (i = 1, 2, 3) \tag{5}$$

点 $\mathbf{x}$ から 3 つのカーネルまでの距離 d_i $(i = 1, 2, 3)$ は次の通りである．（導出は AppendixA を参照）

$$d_1 = \left\| \frac{\mathbf{x}_1 + \mathbf{x}_4}{2} - \frac{\mathbf{x}_2 + \mathbf{x}_3}{2} \right\| \tag{6}$$

$$d_2 = \left\| \frac{\mathbf{x}_1 + \mathbf{x}_3}{2} - \frac{\mathbf{x}_2 + \mathbf{x}_4}{2} \right\| \tag{7}$$

$$d_3 = \left\| \frac{\mathbf{x}_1 + \mathbf{x}_2}{2} - \frac{\mathbf{x}_3 + \mathbf{x}_4}{2} \right\| \tag{8}$$

このうち最短距離を与えるものを平行四辺形までの距離 d とする．幾何学的には 2 組の対角線の中点の間の距離を意味しており，中点が一致するとき距離はゼロとなり図形は平行四辺形となる．

$$d = \min\{d_1, d_2, d_3\} \tag{9}$$

2.5 平行四辺形距離の正規化

合同・相似な図形には同じ平行四辺形距離を付与したい．平行移動，拡大（縮小），回転に対して不変となるよう正規化するための正規化項 Z は次の通りである[5](AppendixB)

$$Z = \left(\sum_{i=1}^{4} \|\mathbf{x}_i - \mathbf{m}\|^2 \right)^{1/2} \quad \text{where } \mathbf{m} := \frac{1}{4} \sum_{i=1}^{4} \mathbf{x}_i \tag{10}$$

3)　vec オペレータは行列を列ベクトルに並べ替える作用素．
4)　t は転置，$\otimes$ はクロネッカー積，I_d は d 次元の単位行列．
5)　幾何学的には，4 つのベクトルをその重心が原点となるよう平行移動して（中心化して）$4d$ 次元空間におけるノルムを正規化項としている．

3　実験手法

3.1　目的とアプローチ

実コーパスから生成された単語ベクトルにおいて類推関係にある単語群を判別できるか平行四辺形距離を用いて評価する．類推関係にある 4 つ組の単語群を 1 群，無作為に抽出した 4 つ組を 2 群として，それぞれについて算出された正規化済みの平行四辺形距離の分布について t 検定を行う．

3.2　単語ベクトル生成のためのデータ

単語共起カウントによる単語ベクトル　実コーパスより単語共起頻度をカウントして単語ベクトルを生成する．English wikipedia dump(20171001 時点）を使用した．トークン数は約 79 億語，語彙数は約 260 万語，前後 5 語以内に生起した単語を共起単語としてカウントする．テストセットの語彙との共起頻度上位 10000 語と，テストセットの語彙の和集合になる計 10072 語の単語共起行列について，その成分を対数化したものを `logfreq10000` とする [7]．さらに，`logfreq10000` に対して特異値分解を行い 300 次元へ次元削減したものを `svd10000` とする．

`logfreq10000` と `svd10000` の行ベクトルをそれぞれの行に対応する単語の単語ベクトルとする

Word2vec　また，ニューラルモデルにより生成された単語ベクトルとして wor2vec[8] の学習済みモデル (spaCy en-core-web-lg[9]) より `logfreq10000` と同じ語彙の単語ベクトルを抽出して `w2v10000` とする．

3.3　正解テストセット

類推関係にある単語群として，Google Analogy Test Set[1] を用いる．これには 19,544 組の単語 4 つ組（例：king-queen-man-woman）が含まれており，14 のカテゴリー（うち 9 つは統語的な類推関係，5 つは意味的な類推関係）に分類されている．

単語 4 つ組は，2 つの単語ペアの組み合わせ (例：king-queen と man-woman) からなっており，ユニークな単語ペアとしては 573 組を含む．Google Analogy Test Set では，全ての単語ペアの組み合わせを網羅していないため，あらためて各カテゴリーごとに全ての単語ペアを組み合わせて 15,247 組の 4 つ組を得た（「正解テストセット」という）．

4　実験結果

4.1　2 群間の比較

図 2 に，類推関係を持つ単語群 (“analogy_testset”) と無作為に抽出した 4 単語 (“random_sampling”) について算出された平行四辺形距離の分布を示す．

`logfreq10000` では類推関係の単語群の図形距離は平均 0.321 であるのに対して，無作為抽出は平均 0.436 であり，その差は 0.115 で統計的に有意な違いが見られた ($t(15246) = 114.7, p < 0.001$)．同様に `svd10000` では analogy-testset の平均値 0.322 と random_sampling の平均値 0.435 の差は 0.113($t(15246) = 112.1, p < 0.001$) と有意である．

また，`w2v10000` においても analogy_testset の平均値 0.391 に対し random_sampling の平均値 0.535 と類推関係にある単語群の距離は 0.144 小さく有意な差を示した ($t(8576) = 145.6, p < 0.001$) が，その分布形状の差は著しい．なお，`w2v10000` に欠けている語彙を含むカテゴリー capital-word を除外したためサンプル数が異なる．

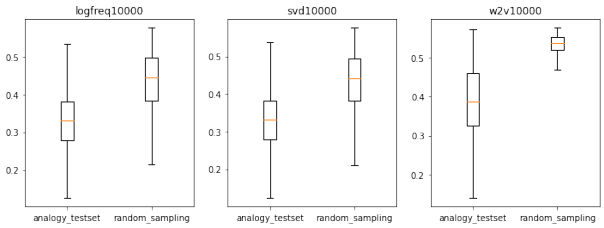

図 2　図形距離の比較

4.2　分析

カテゴリーごとに比較すると，最も平行四辺形に近いカテゴリーは family（平均 0.257），次いで gram4-superlative(平均 0.277）であった．反対に平行四辺形から遠いカテゴリーは gram1-adjective-to-adverb（平均 0.369），次いで opposite(平均 0.358）であった．

平均図形距離が最小の family についてカテゴリー内の順位を見ると，最も平行四辺形に近い単語群は “his-her-nephew-niece” の 4 つ組であり，その図形距離 0.095 は全カテゴリーの中でも最小であった．反対に，最も平行四辺形から遠い単語群は，“policeman-policewoman-stepbrother-stepsister” であり，その平行四辺形距離は 0.549 である．無作為抽出を含めて算出した値の中での最大値に近い値である．

t 検定の結果は平行四辺形距離が類推関係にある単語群を有意に特徴づけることを示しているが，そ

表 1 頂点の配置パタン占率 (%)(logfreq10000)

カテゴリー	パタン 1	パタン 2	パタン 3
capital-common-countries	93.28	3.95	2.77
capital-world	96.48	1.53	1.99
city-in-state	54.92	2.90	42.19
currency	69.43	11.72	18.85
family	72.73	26.09	1.19
gram1-adjective-to-adverb	53.63	39.72	6.65
gram2-opposite	72.41	9.11	18.47
gram3-comparative	91.89	0.00	8.11
gram4-superlative	89.66	5.53	4.81
gram5-present-participle	60.42	34.66	4.92
gram6-nationality-adjective	95.61	2.56	1.83
gram7-past-tense	55.13	43.85	1.03
gram8-plural	87.69	12.16	0.15
gram9-plural-verbs	81.38	14.48	4.14

の順位を見ると低頻度の単語を含む場合に特に平行四辺形距離のばらつきが大きく，識別精度に課題がある．本論文では紙面の都合上割愛するが正規化項を通じてノルムが過大な影響を及ぼしていると思われ，今後の研究課題である．

5 考察

5.1 ベクトルの逆転現象

式 (9) により平行四辺形距離を算出する際に，最小距離を与える頂点配置のパタンを各カテゴリーごとに示したのが表 1 である．

正解テストセットに含まれる単語ペアは関係の方向を反映した並びとなっており，例えば family カテゴリーでは father-mother のように男性・女性の並びである．このような単語ペアを組み合わせて 4 つ組を作ると，単語ベクトル $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ はペア 1 の男性，女性，ペア 2 の男性，女性という並びとなる．従って類推関係が成立していれば $\mathbf{x}_1 - \mathbf{x}_2 \simeq \mathbf{x}_3 - \mathbf{x}_4$ が成り立って，3 つのカーネルまでの距離のうち $d_1 \simeq 0$ が最短距離となる (パタン 1 に対応)．

しかるに，family など 4 つのカテゴリーでパタン 2 が 2 割以上を占める．つまり式 (7) が適用され $\mathbf{x}_1 - \mathbf{x}_2 \simeq \mathbf{x}_4 - \mathbf{x}_3$ が成り立っていることになる．これが示唆するのは，男性単語と女性単語の間の差ベクトルの向きが 2 つのペア間で逆転する現象が起きているということである．

表 2 に差ベクトルの方向に対応するペアを例示する．順方向のペアを組み合わせた 4 つ組の場合，例えば $\boldsymbol{v}_{he} - \boldsymbol{v}_{she} = \boldsymbol{v}_{father} - \boldsymbol{v}_{mother}$ が成り立つので $\boldsymbol{v}_{he} - \boldsymbol{v}_{she} + \boldsymbol{v}_{mother} = \boldsymbol{v}_{father}$ の演算により左辺の 3 単語から father を推定できるが，逆転が起きているペアの場合 $\boldsymbol{v}_{he} - \boldsymbol{v}_{she} = \boldsymbol{v}_{wife} - \boldsymbol{v}_{husband}$ が成立しているので，四項類推演算を行う場合には $\boldsymbol{v}_{he} - \boldsymbol{v}_{she} + \boldsymbol{v}_{wife} = \boldsymbol{v}_{husband}$ ではなく $\boldsymbol{v}_{he} - \boldsymbol{v}_{she} + \boldsymbol{v}_{husband} = \boldsymbol{v}_{wife}$ の演算が成り立つ．

表 2 順方向と逆方向の差分ベクトルを持つ単語ペア事例

カテゴリー	順方向	逆方向
family	he-she father-mother	husband-wife groom-bride
adj-to-adv	free-freely calm-calmly	immediate-immediately quick-quickly
present-participle	think-thinking generate-generating	sing-singing swim-swimming

5.2 逆転現象の要因

単語ペアの差ベクトルが反転する原因は，単語ベクトルのノルムの逆転が起きているからである．例えば family カテゴリーの多くのペアでは男性単語ベクトルのノルムは女性単語のそれよりも大きいが，反転しているペアの場合に女性単語のノルムの方が大きい．言語現象上の要因としては，コーパス中の出現頻度において男女バイアスが存在しているからであり，一般に男性単語の方が頻度が高いが，逆転しているペアでは，例えば husband よりも wife が頻出するのは his wife のように男性の文脈において wife が出現することの方が her husband よりも多いということが想定される．

同じカテゴリーの中に異なる関係性を持つペアが存在してサブカテゴリーをなしていることを示唆しており，平行四辺形距離はそのような単語間の関係性を定量的に測定する機能を有している．

6 おわりに

本研究では単語群の持つ意味的・統語的関係を幾何的性質として捉え，新たなメトリックとして図形距離を提案した．そして類推関係にある単語群が高次元空間内で平行四辺形に近い配置にあることを実証した．精度の改善のため正規化手法の見直しが必要であるが，図形距離を用いることで，半自動的な類推関係の発見や，類推関係の種類の細分化などが期待される．また図形距離は本稿で対象とした平行四辺形以外の図形にも拡張できる．そのため他の立体図形をなす単語群の研究も進行中である．[6)]

6) 例えば，正四面体をなす単語群として spring-summer-autumn-winter がある．

謝辞

本研究は JST さきがけ JPMJPR20C9 の助成を受けて行われた.

参考文献

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeff Dean. Efficient estimation of word representations in vector space. pp. 1310–4546, 2013.

[2] Dawn Chen, Joshua C. Peterson, and Thomas L. Griffiths. Evaluating vector-space models of analogy. **CoRR**, Vol. abs/1705.04416, , 2017.

[3] Joshua C. Peterson, Dawn Chen, and Thomas L. Griffiths. Parallelograms revisited: Exploring the limitations of vector space models for simple analogies. **Cognition**, Vol. 205, p. 104440, 2020.

[4] 丸山圭三郎. ソシュールを読む. 岩波書店, 1983.

[5] 河田敬義. アフィン幾何・射影幾何. 岩波書店, 1967.

[6] 前田晃弘, 鳥居拓馬, 日髙昇平. 単語共起行列の内部構造解明のための構成論的アプローチ. 2022 年度日本認知科学会第 39 回大会, 2022.

[7] Takuma Torii, Akihiro Maeda, and Shohei Hidaka. Embedding parallelepiped in co-occurrence matrix: simulation and empirical evidence. In **Joint Conference on Language Evolution (JCoLE2022)**, 2022.

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. pp. 1310–4546, 2013.

[9] spaCy. Models and language, 2023. https:spacy.io/usage/models.

A　カーネルまでの距離の導出

平行四辺形の必要十分条件を示す式 (3) は，次のようにベクトル化することができる．

$$\mathrm{vec}(X\mathbf{p}_i) = (\mathbf{p}_i^t \otimes I_d)\mathrm{vec}(X) = A_i\mathbf{x}$$

ただし，$A_i := \mathbf{p}_i^t \otimes I_d$, $\mathrm{vec}(X) = \mathbf{x} \in \mathbb{R}^{4d}$ である．
平行四辺形をなすすべての点の集合は A_i のカーネル K_i として次の通りである．

$$K_i = \{\mathbf{x} \in \mathbb{R}^{4d} | A_i\mathbf{x} = \mathbf{0}_d\}$$

頂点パタン i の平行四辺形までの距離を点 $\mathbf{x}$ からカーネル K_i までの距離として求める．K_i は行列 A_i の行空間 $C(A_i)$ の補空間であることを踏まえ ($\mathbb{R}^{4d} = K_i \oplus C(A_i)$)，まず $C(A_i)$ への射影行列 Π_i を考える．$\mathbf{x}$ を $C(A_i)$ へ直交射影した点を $\mathbf{z}_i$ とする．すなわち，$\Pi_i\mathbf{x} = \mathbf{z}_i$ である．$\mathbf{z}_i \in C(A_i)$ であるので，$A_i^t\mathbf{w} = \mathbf{z}_i$ なる $\mathbf{w}$ が存在する．

点 $\mathbf{x}$ と点 $\mathbf{z}_i$ を結ぶ直線は部分空間 $C(A_i)$ への垂線となるので $A_i(\mathbf{x} - \mathbf{z}_i) = \mathbf{0}$ が成り立ち，これを展開すると $\mathbf{w} = (A_iA_i^t)^{-1}A_i\mathbf{x}$ を得て次の式に代入する．

$$\begin{aligned}\Pi_i\mathbf{x} &= \mathbf{z}_i \\ &= A_i^t\mathbf{w} \\ &= A_i^t(A_iA_i^t)^{-1}A_i\mathbf{x}\end{aligned}$$

この式は任意の $\mathbf{x}$ について成り立つので $\Pi_i = A_i^t(A_iA_i^t)^{-1}A_i$ である．$A_i = \mathbf{p}_i^t \otimes I_d$ を用いて，さらに式を展開する．$i = 1$ の場合，

$$\begin{aligned}\Pi_1\mathbf{x} &= A_1^t(A_1A_1^t)^{-1}A_1\mathbf{x} \\ &= \Big(\frac{1}{4}(\mathbf{p}_1\mathbf{p}_1^t) \otimes I_d\Big)\mathbf{x} \\ &= \frac{1}{4}\begin{bmatrix} I_d & -I_d & -I_d & I_d \\ -I_d & I_d & I_d & -I_d \\ -I_d & I_d & I_d & -I_d \\ I_d & -I_d & -I_d & I_d \end{bmatrix}\begin{bmatrix}\mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \mathbf{x}_4\end{bmatrix} \\ &= \frac{1}{4}\begin{bmatrix}\mathbf{x}_1 - \mathbf{x}_2 - \mathbf{x}_3 + \mathbf{x}_4 \\ -\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3 - \mathbf{x}_4 \\ -\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3 - \mathbf{x}_4 \\ \mathbf{x}_1 - \mathbf{x}_2 - \mathbf{x}_3 + \mathbf{x}_4\end{bmatrix}\end{aligned}$$

となるので

$$d_1 = ||\Pi_1\mathbf{x}|| = \frac{1}{2}||X\mathbf{p}_1|| = \left\|\frac{\mathbf{x}_1 + \mathbf{x}_4}{2} - \frac{\mathbf{x}_2 + \mathbf{x}_3}{2}\right\| \quad (11)$$

である．$i = 2, 3$ の場合も同様である．

B　正規化項の導出

点 $\mathbf{x}$ からそれをカーネル K_i へ直交射影した点までの垂線の距離を，原点 O から点 $\mathbf{x}$ までの斜辺の距離で正規化すれば正弦が得られるが，その方法では図形の平行移動の影響を受けてしまう．これを避けるために，空間 $V = \mathbb{R}^d$ における 4 点を任意のベクトル $\mathbf{b} \in V$ で平行移動したものと同一視する．これは空間 $W = \mathbb{R}^{4d}$ においては，$\mathbf{x} \in W$ と $\mathbf{x} + \mathbf{1}_4 \otimes \mathbf{b} \in W$ を同一視することに等しい．

任意の点 $\mathbf{x}$ に対して，原点と同一視できる点の中で最短距離を与えるものをその原点して選ぶこととして，平行移動 $\mathbf{b}' = \mathbf{1}_4 \otimes \mathbf{b}$ を次のように求める．

$$\begin{aligned}\min_{\mathbf{b}'}\delta &= \min_{\mathbf{b}'}\left\|\mathbf{x} - \mathbf{b}'\right\| \\ &= \min_{\mathbf{b}}\left\|\mathbf{x} - \mathbf{1}_4 \otimes \mathbf{b}\right\| \\ &= \min_{\mathbf{b}}\Big(\sum_{i=1}^{4}(\mathbf{x}_i - \mathbf{b})^2\Big)^{1/2}\end{aligned}$$

δ^2 を $\mathbf{b}$ で偏微分する．

$$\frac{\partial\delta^2}{\partial\mathbf{b}} = -2\sum_{i=1}^{4}\mathbf{x}_i + 8\mathbf{b} = 0$$

従って，

$$\mathbf{b} = \frac{1}{4}\sum_{i=1}^{4}\mathbf{x}_i := \mathbf{m}$$

すなわち，空間 V における 4 つの単語ベクトルの中点を $\mathbf{m}$ とすると，空間 W においてこれに対応する $\mathbf{1}_4 \otimes \mathbf{m}$ が $\mathbf{x}$ に最も近い原点（平行移動により原点と同一視される点) ということになる．

以上より，図形距離の正規化に用いる正規化項 Z は，$\mathbf{1}_4 \otimes \mathbf{m}$ までの距離ということになる．

$$\begin{aligned}Z &= \|\mathbf{x} - \mathbf{1}_4 \otimes \mathbf{m}\| \\ &= \Big(\sum_{i=1}^{4}\|\mathbf{x}_i - \mathbf{m}\|^2\Big)^{1/2} \\ &= \Big(\sum_{i=1}^{4}\|\overline{\mathbf{x}_i}\|^2\Big)^{1/2}\end{aligned}$$

$\overline{\mathbf{x}_i}$ は，4 点の重心 $\mathbf{m}$ で中心化したベクトルと考えることができるので，これは結局，4 つのベクトルを重心の分だけマイナスに平行移動していることであり，合同となる図形を同一視して，その正規化項を導出していることになる．