"Metode de optimizare Riemanniene pentru învățare profundă"
Proiect cofinanțat din Fondul European de Dezvoltare Regională prin
Programul Operațional Competitivitate 2014-2020

On The Information Geometry of Word Embedding

Riccardo Volpi, joint work with D. Marinelli, P. Hlihor, and L. Malagò

Romanian Institute of Science and Technology

Synergies in GDA Workshop 08 December, 2017

Word Embedding

A word embedding maps the words of a dictionary in a real vector space, based on the notion of context

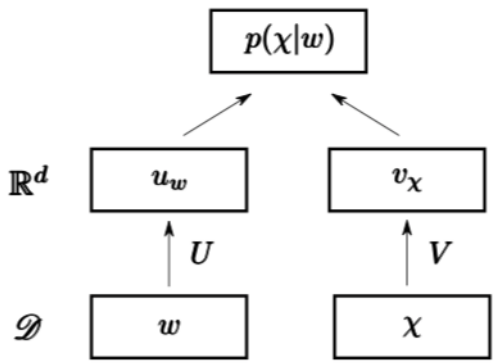
> *"You shall know a word by the company it keeps"*. Firth, 1957.

Word Embedding

A word embedding maps the words of a dictionary in a real vector space, based on the notion of context

"You shall know a word by the company it keeps". Firth, 1957.

$$p(\chi|w) = \exp(u_w^{\mathrm{T}} v_\chi)/Z_w$$



- The general model used by Skip-Gram (Mikolov et. al., '13) and Glove (Pennington et. al., '14)

Analogies of the form $a : b = c : d$ can be solved by

$$\arg\min_d \ \|u_a - u_b - u_c + u_d\|^2 =$$

$$= \arg\min_c \sum_{\chi \in \mathscr{D}} \left(\ln \frac{p(\chi|a)}{p(\chi|b)} - \ln \frac{p(\chi|c)}{p(\chi|d)} \right)^2$$

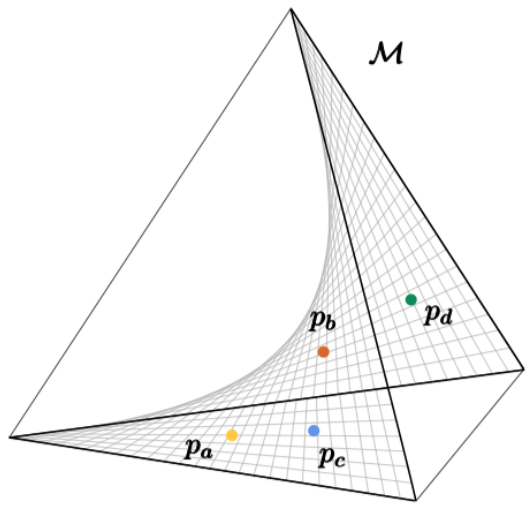
- The space of word embedding has a linear geometry (cf. Arora et. al., '16), where vectors express semantic relationships between contexts

Exponential Family and Conditional Distributions

Consider the joint probability distribution for W and $\mathcal{X}$

$$p(\chi, w) = \exp(w^{\mathrm{T}} C \chi)/Z, \qquad \text{with } C = U^{\mathrm{T}} V$$

- Conditional distributions $p(\chi|w) = \exp(u_w^{\mathrm{T}} v_\chi)/Z_w$ lay on the boundary of the joint statistical model
- Each column vector of U identifies a p_w in the conditional model
- For a fixed V, all conditional simplexes are homomorphic one to each other

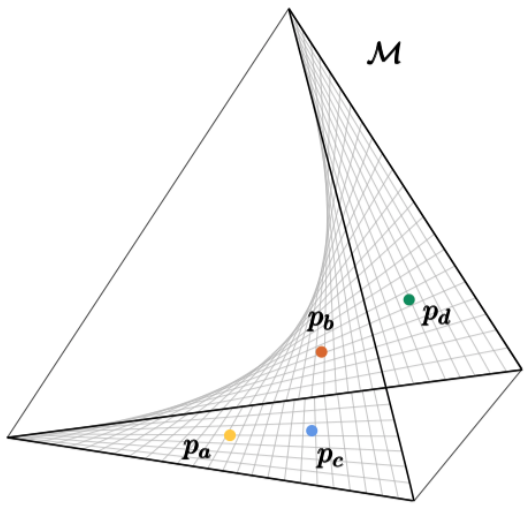


Exponential Family and Conditional Distributions

Consider the joint probability distribution for W and $\mathcal{X}$

$$p(\chi, w) = \exp(w^{\mathrm{T}} C \chi)/Z, \qquad \text{with } C = U^{\mathrm{T}} V$$

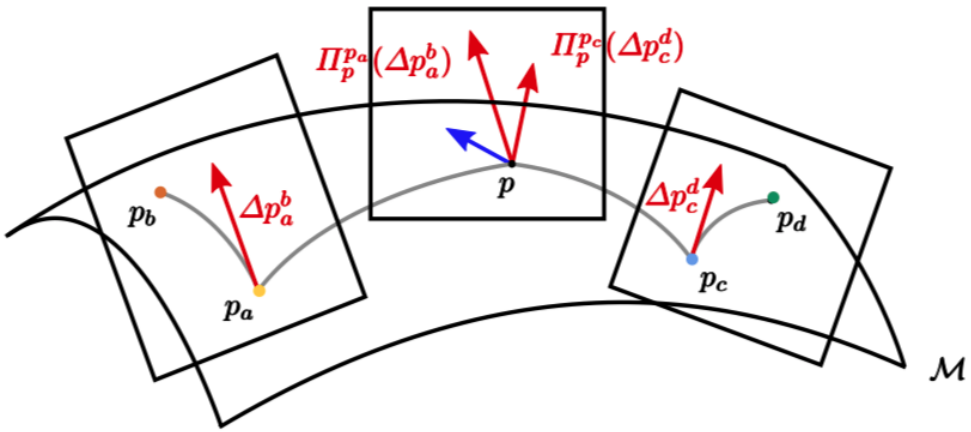
- Conditional distributions $p(\chi|w) = \exp(u_w^{\mathrm{T}} v_\chi)/Z_w$ lay on the boundary of the joint statistical model
- Each column vector of U identifies a p_w in the conditional model
- For a fixed V, all conditional simplexes are homomorphic one to each other



We aim at characterizing the geometry of word embedding, based on alternative geometries for the exponential family studied in Information Geometry (Amari and Nagaoka, '00)

Geometric Word Analogies

Let p_w be the conditional probability $p(\chi|W = w)$, and p a reference context



- The logarithmic map $\mathcal{M} \to \mathsf{T}_p\mathcal{M}$ is defined by

$$\Delta p_a^b = \mathrm{Log}_{p_a}(p_b)$$

- The parallel transport of A

$$\Pi_p^{p_a} A : \mathsf{T}_{p_a}\mathcal{M} \to \mathsf{T}_p\mathcal{M}$$

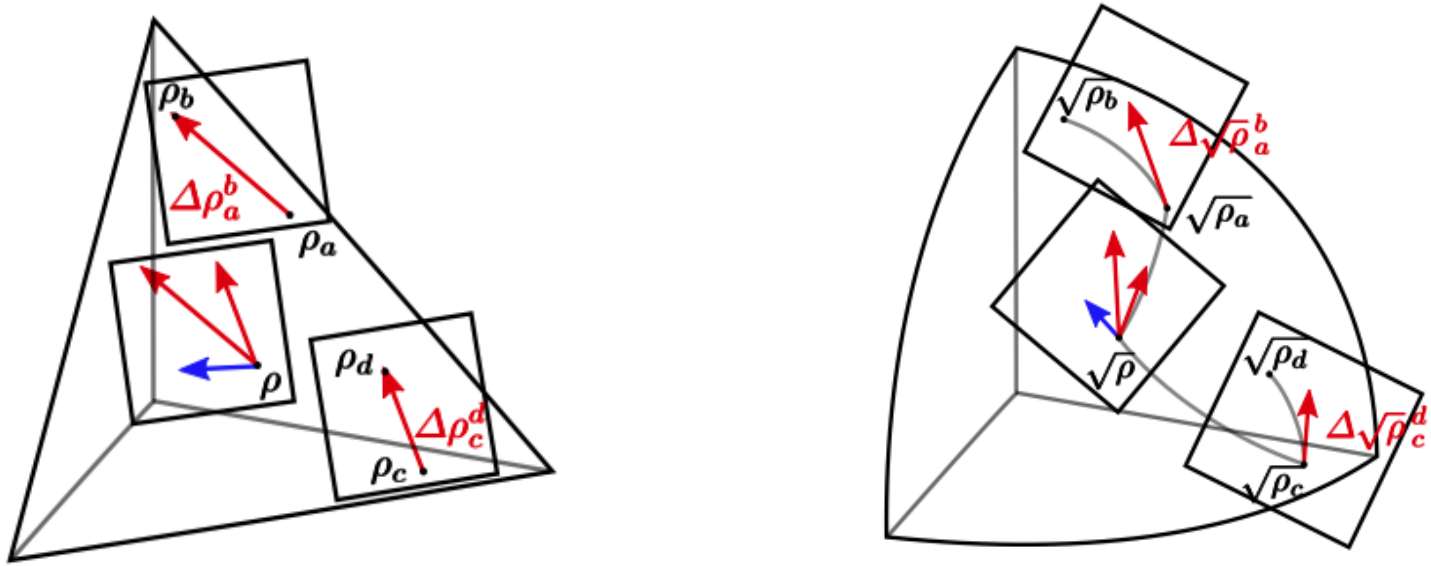
- Norms are computed by $\|A\|_p^2 = a^{\mathrm{T}} I(p) a$, where $I(p)$ is the Fisher information matrix

Analogies of the form $a : b = c : ?$ can solved by

$$\arg\min_d \left\| \Pi_p^{p_a} \Delta p_a^b - \Pi_p^{p_c} \Delta p_c^d \right\|_p^2 ,$$

The Framework in Practice: The Full Simplex

- For $d = \#(\mathscr{D})$, any point $(\rho)_\chi$ in the interior of the simplex corresponds to a conditional probability $p(\chi|W = w)$
- By setting $\rho \mapsto \sqrt{\rho}$, the probability simplex is mapped to the positive spherical orthant and the geometry of the sphere is obtained



The Framework in Practice: The Exponential Family

- For $d \leq \#(\mathscr{D})$, the Riemannian geometry of the exponential family is defined by the Fisher-Rao metric
- Moreover, there are at least two other affine geometries of interest: the exponential geometry and the mixture geometry

The Framework in Practice: The Exponential Family

- For $d \leq \#(\mathscr{D})$, the Riemannian geometry of the exponential family is defined by the Fisher-Rao metric
- Moreover, there are at least two other affine geometries of interest: the exponential geometry and the mixture geometry

- [Proposition] Let p_0 be the uniform distribution over $\mathscr{D}$, ${}^{e}\Pi_{p}^{q}$, and ${}^{e}\Delta p_{a}^{b}$ be defined according to the exponential geometry, under the hypothesis of isotropy distribution for the v's

$$\arg\min_{d} \left\| {}^{e}\Pi_{p}^{p_a}({}^{e}\Delta p_{a}^{b}) - {}^{e}\Pi_{p}^{p_c}({}^{e}\Delta p_{c}^{d}) \right\|_{p_0}^{2} ,$$

reduces to

$$\arg\min_{d} \left\| u_a - u_b - u_c + u_d \right\|^{2} ,$$

Conclusions and Future Perspectives

- The language of Information Geometry can be used to describe the geometry of word embeddings
- We have defined a parameter-invariant way to solve word analogies
- The exponential geometry of the exponential family allows to recover the standard way to solve word analogies
- Evaluating experimentally the role of different geometries of word embedding

Conclusions and Future Perspectives

- The language of Information Geometry can be used to describe the geometry of word embeddings
- We have defined a parameter-invariant way to solve word analogies
- The exponential geometry of the exponential family allows to recover the standard way to solve word analogies
- Evaluating experimentally the role of different geometries of word embedding

"One geometry cannot be more true than another; it can only be more convenient". Henri Poincaré, Science and Hypothesis, 1902.