# Master M2 MVA 2018/2019
# Deep Learning - MP2

Souhaib ATTAIKI

December 29, 2018

## 2 - Multilingual word embeddings

We want to prove that :

$$W^\star = \underset{W \in \mathcal{O}_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_F = UV^T$$

with $U\Sigma V^T = SVD(YX^T)$.

This is equivalent to :

$$\begin{aligned}
W^\star &= \underset{W \in \mathcal{O}_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_F^2 \\
&= \underset{W \in \mathcal{O}_d(\mathbb{R})}{\operatorname{argmin}} \left(\|WX\|_F^2 + \|Y\|_F^2 - 2\langle WX, Y\rangle\right) \\
&= \underset{W \in \mathcal{O}_d(\mathbb{R})}{\operatorname{argmin}} \left(\|X\|_F^2 + \|Y\|_F^2 - 2Tr(X^T W^T Y)\right) \\
&= \underset{W \in \mathcal{O}_d(\mathbb{R})}{\operatorname{argmax}} Tr(X^T W^T Y) \\
&= \underset{W \in \mathcal{O}_d(\mathbb{R})}{\operatorname{argmax}} Tr(YX^T W^T)
\end{aligned}$$

Where we used the fact that $\|W\|_F^2 = 1$ and that $\|X\|_F^2 + \|Y\|_F^2$ does not depend on $W$.

However, we have $YX^T = U\Sigma V$ and $W = U_w \Sigma_w V_w^T$ ($W \in \mathcal{O}_d(\mathbb{R})$), so :

$$\begin{aligned}
Tr(YX^T W^T) &= Tr(U\Sigma V^T V_w \Sigma_w U_w^T) \\
&= Tr(\widetilde{U}\Sigma\widetilde{V}\Sigma_w) \quad \text{with } \widetilde{U} = U_w^T U \text{ and } \widetilde{V} = V^T V_w \\
&\leqslant Tr(\Sigma\Sigma_w) \quad \text{by applying } \textit{Von Neumann's trace inequality}
\end{aligned}$$

So the optimal value is such that $\widetilde{U} = \mathbb{1}$ and $\widetilde{V} = \mathbb{1}$. Thereby, $U = U_w$ and $V = V_w$, so since $W = U\Sigma_w V^T \in \mathcal{O}_d(\mathbb{R})$, we have $U\Sigma_w^2 U^T = \mathbb{1}$ so $\Sigma_w^2 = \Sigma_w = \mathbb{1}$, which leads us to $W = UV^T$.

# 3 - Sentence classification with BoV

See *notebook*.

# 4 - Deep Learning models for classification

**Which loss did you use ? Write the mathematical expression of the loss you used for the 5-class classification** I've used the *'categorical_crossentropy'* loss. The mathematical expression of the loss is :

$$L = \frac{1}{n_{obs}} \sum_{i=1}^{n_{obs}} \sum_{c=1}^{5} \mathbb{1}_{y_i \in C_c} \log(p_{model}(y_i \in C_c))$$

where $y_i$ is the output of the model, $C_c$ is the category $c$, and $p_{model}(y_i \in C_c)$ is the probability predicted by the model for the 'i'th observation to belong to the $c$ th category.

**Plot the evolution of train/dev results w.r.t the number of epochs** See *notebook*.