

Supplementary Material

Equivalent formulations of $\mathbf{T}^{\mathbf{AB}}$

$\mathbf{T}^{\mathbf{AB}}$ is defined as the unique PSD matrix verifying $\mathbf{T}^{\mathbf{AB}} \mathbf{A} \mathbf{T}^{\mathbf{AB}} = \mathbf{B}$. Using this definition, we derive two equivalent formulations for $\mathbf{T}^{\mathbf{AB}}$:

$$\begin{aligned}\mathbf{T}^{\mathbf{AB}} &= \mathbf{A}^{-\frac{1}{2}} \left(\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}} \right)^{\frac{1}{2}} \mathbf{A}^{-\frac{1}{2}} \\ &= \mathbf{B}^{\frac{1}{2}} \left(\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}} \right)^{-\frac{1}{2}} \mathbf{B}^{\frac{1}{2}}\end{aligned}$$

The first is derived as in [Malagò et al., 2018]:

$$\begin{aligned}\mathbf{T}^{\mathbf{AB}} \mathbf{A} \mathbf{T}^{\mathbf{AB}} &= \mathbf{B} \\ \mathbf{A}^{\frac{1}{2}} \mathbf{T}^{\mathbf{AB}} \mathbf{A}^{\frac{1}{2}} \mathbf{A}^{\frac{1}{2}} \mathbf{T}^{\mathbf{AB}} \mathbf{A}^{\frac{1}{2}} &= \mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}} \\ \mathbf{A}^{\frac{1}{2}} \mathbf{T}^{\mathbf{AB}} \mathbf{A}^{\frac{1}{2}} &= \left(\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}} \right)^{\frac{1}{2}} \\ \mathbf{T}^{\mathbf{AB}} &= \mathbf{A}^{-\frac{1}{2}} \left(\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}} \right)^{\frac{1}{2}} \mathbf{A}^{-\frac{1}{2}}\end{aligned}$$

We then adapt this derivation to obtain a second formulation of $\mathbf{T}^{\mathbf{AB}}$:

$$\begin{aligned}\mathbf{T}^{\mathbf{AB}} \mathbf{A} \mathbf{T}^{\mathbf{AB}} &= \mathbf{B} \\ (\mathbf{T}^{\mathbf{AB}})^{-1} \mathbf{B} (\mathbf{T}^{\mathbf{AB}})^{-1} &= \mathbf{A} \\ \mathbf{B}^{\frac{1}{2}} (\mathbf{T}^{\mathbf{AB}})^{-1} \mathbf{B}^{\frac{1}{2}} \mathbf{B}^{\frac{1}{2}} (\mathbf{T}^{\mathbf{AB}})^{-1} \mathbf{B}^{\frac{1}{2}} &= \mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}} \\ \mathbf{B}^{\frac{1}{2}} (\mathbf{T}^{\mathbf{AB}})^{-1} \mathbf{B}^{\frac{1}{2}} &= \left(\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}} \right)^{\frac{1}{2}} \\ (\mathbf{T}^{\mathbf{AB}})^{-1} &= \mathbf{B}^{-\frac{1}{2}} \left(\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}} \right)^{\frac{1}{2}} \mathbf{B}^{-\frac{1}{2}} \\ \mathbf{T}^{\mathbf{AB}} &= \mathbf{B}^{\frac{1}{2}} \left(\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}} \right)^{-\frac{1}{2}} \mathbf{B}^{\frac{1}{2}}\end{aligned}$$

Derivation of the Riemannian gradient updates

From [Malagò et al., 2018], we have that the exp and log maps of the Riemannian Bures metric are given by:

$$\begin{aligned}\exp_{\mathbf{C}}(\mathbf{V}) &= (\mathcal{L}_{\mathbf{C}}(\mathbf{V}) + \mathbf{I}) \mathbf{C} (\mathcal{L}_{\mathbf{C}}(\mathbf{V}) + \mathbf{I}) \\ \log_{\mathbf{C}}(\mathbf{B}) &= (\mathbf{T}^{\mathbf{CB}} - \mathbf{I}) \mathbf{C} + \mathbf{C} (\mathbf{T}^{\mathbf{CB}} - \mathbf{I})\end{aligned}$$

where $\mathcal{L}_{\mathbf{C}}(\mathbf{V})$ is the solution of *Lyapunov* equation $\mathcal{L}_{\mathbf{C}}(\mathbf{V}) \mathbf{C} + \mathbf{C} \mathcal{L}_{\mathbf{C}}(\mathbf{V}) = \mathbf{V}$. One can show that the $\mathcal{L}_{\mathbf{C}}$ operator is linear, and that the following identity holds: $\mathcal{L}_{\mathbf{C}}(\mathbf{X} \mathbf{C} + \mathbf{C} \mathbf{X})$. In particular, $\mathcal{L}_{\mathbf{C}}(\log_{\mathbf{C}} \mathbf{B}) = \mathbf{T}^{\mathbf{CB}} - \mathbf{I}$.

From this, since $\text{grad}_{\mathbf{A}} \frac{1}{2} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = -\log_{\mathbf{A}} \mathbf{B}$, the Riemannian gradient update is given by

$$\begin{aligned}\mathbf{A}_{t+1} &= \exp_{\mathbf{A}_t}(\eta_t \log_{\mathbf{A}_t} \mathbf{B}) \\ &= (\eta_t \mathcal{L}_{\mathbf{A}_t}(\log_{\mathbf{A}_t} \mathbf{B}) + \mathbf{I}) \mathbf{A}_t (\eta_t \mathcal{L}_{\mathbf{A}_t}(\log_{\mathbf{A}_t} \mathbf{B}) + \mathbf{I}) \\ &= ((1 - \eta_t)\mathbf{I} + \eta_t \mathbf{T}^{\mathbf{A}_t \mathbf{B}}) \mathbf{A}_t ((1 - \eta_t)\mathbf{I} + \eta_t \mathbf{T}^{\mathbf{A}_t \mathbf{B}})\end{aligned}$$

Derivation of the Euclidean gradient

Notations: \otimes is the Kronecker product of matrices. Recall that

$$\begin{aligned}[\mathbf{B}^\top \otimes \mathbf{A}] \text{vec}(\mathbf{X}) &= \text{vec}(\mathbf{A} \mathbf{X} \mathbf{B}) \\ [\mathbf{A} \otimes \mathbf{B}][\mathbf{C} \otimes \mathbf{D}] &= [\mathbf{A} \mathbf{C} \otimes \mathbf{B} \mathbf{D}]\end{aligned}$$

In the following, we will often omit the $\text{vec}(\cdot)$ and treat matrices as vectors when the context makes it clear. We will make use of the following identities:

$$\begin{aligned}\partial_{\mathbf{X}} f \circ g(\mathbf{X}) &= \partial_{\mathbf{X}} f(g(\mathbf{X})) \partial_{\mathbf{X}} g(\mathbf{X}) \\ \partial_{\mathbf{X}} (fg)(\mathbf{X}) &= [g(\mathbf{X})^\top \otimes \mathbf{I}] \partial_{\mathbf{X}} f(\mathbf{X}) + [\mathbf{I} \otimes g(\mathbf{X})] \partial_{\mathbf{X}} g(\mathbf{X})\end{aligned}$$

and

$$\partial_{\mathbf{X}} \mathbf{X}^{\frac{1}{2}} = [\mathbf{X}^{\frac{1}{2}} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{X}^{\frac{1}{2}}]^{-1}$$

Let $f(\mathbf{A}, \mathbf{B}) = \text{Tr}(\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}})^{\frac{1}{2}}$.

Let us differentiate f w.r.t \mathbf{A} :

$$\begin{aligned}\nabla_{\mathbf{A}} f(\mathbf{A}, \mathbf{B}) &= \left[\partial_{\mathbf{A}} (\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}})^{\frac{1}{2}} \right]^\top \mathbf{I} \\ &= \left[\left[\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}} \right]^{\frac{1}{2}} \otimes \mathbf{I} + \mathbf{I} \otimes (\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}})^{\frac{1}{2}} \right]^{-1} \partial_{\mathbf{A}} (\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}}) \right]^\top \mathbf{I} \\ &= \left[\mathbf{B}^{\frac{1}{2}} \otimes \mathbf{B}^{\frac{1}{2}} \right] \left[(\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}})^{\frac{1}{2}} \otimes \mathbf{I} + \mathbf{I} \otimes (\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}})^{\frac{1}{2}} \right]^{-1} \mathbf{I} \\ &= \left[\mathbf{B}^{\frac{1}{2}} \otimes \mathbf{B}^{\frac{1}{2}} \right] \frac{1}{2} (\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}})^{-\frac{1}{2}} \\ &= \frac{1}{2} \mathbf{B}^{\frac{1}{2}} (\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}})^{-\frac{1}{2}} \mathbf{B}^{\frac{1}{2}}\end{aligned}$$

Therefore $\nabla_{\mathbf{A}} f(\mathbf{A}, \mathbf{B}) = \frac{1}{2} \mathbf{T}^{\mathbf{A} \mathbf{B}}$

Let now $\mathbf{A} = \mathbf{L} \mathbf{L}^\top$, let us differentiate w.r.t \mathbf{L} :

$$\begin{aligned}\nabla_{\mathbf{L}} f(\mathbf{L} \mathbf{L}^\top, \mathbf{B}) &= \left[\partial_{\mathbf{L}} (\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}})^{\frac{1}{2}} \right]^\top \mathbf{I} \\ &= \partial_{\mathbf{L}} \mathbf{A}^\top \left[\partial_{\mathbf{A}} (\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}})^{\frac{1}{2}} \right]^\top \mathbf{I} \\ &= [\mathbf{L}^\top \otimes \mathbf{I}] [\mathbf{I} + \mathbf{T}_{n,n}] \frac{1}{2} \mathbf{B}^{\frac{1}{2}} (\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}})^{-\frac{1}{2}} \mathbf{B}^{\frac{1}{2}} \\ &= \mathbf{B}^{\frac{1}{2}} (\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}})^{-\frac{1}{2}} \mathbf{B}^{\frac{1}{2}} \mathbf{L}\end{aligned}$$

where $\mathbf{T}_{n,n}$ is the transposition tensor, such that $\forall \mathbf{X} \in \mathbb{R}^{n \times n}$, $\mathbf{T}_{n,n} \text{vec}(\mathbf{X}) = \text{vec}(\mathbf{X}^\top)$.

Therefore $\nabla_{\mathbf{L}} f(\mathbf{L} \mathbf{L}^\top, \mathbf{B}) = \mathbf{T}^{\mathbf{A} \mathbf{B}} \mathbf{L}$.

Using the same calculations, one can see that if $\mathbf{A} = \mathbf{L} \mathbf{L}^\top + \varepsilon \mathbf{I}$, then we still have

$$\nabla_{\mathbf{L}} f(\mathbf{L} \mathbf{L}^\top + \varepsilon \mathbf{I}, \mathbf{B}) = \mathbf{T}^{\mathbf{A} \mathbf{B}} \mathbf{L}$$

since $\partial_{\mathbf{L}} [\mathbf{L} \mathbf{L}^\top + \varepsilon \mathbf{I}] = \partial_{\mathbf{L}} [\mathbf{L} \mathbf{L}^\top]$

Model Hyperparameters and Training Details

Word Embeddings We train our embeddings on the concatenated ukWaC and WaCkypedia corpora [Baroni et al., 2009], consisting of about 3 billion tokens, on which we keep only the tokens appearing more than 100 times in the text after lowercasing and removal of all punctuation (for a total number of 261583 different words). We optimize 5 epoches using adagrad [Duchi et al., 2011] with $\epsilon = 10^{-8}$ with a learning rate of 0.01. We use a window size of 10 (i.e. positive examples consist of the first 5 preceding and first 5 succeeding words), set the margin to 10, sample one negative context per positive context and, in order to prevent the norms of the embeddings to be too highly correlated with the corresponding word frequencies (see Figure 7), we use two distinct sets of embeddings for the input and context words. In order to use as much parallelization as possible, we use batches of size 10000, but believe that smaller batches would lead to improved performances. We limit matrix square root approximations to 6 Newton-Schulz iterations and add $0.01\mathbf{I}$ to the covariances to ensure non-singularity.

To generate batches, we use the same sampling tricks as in [Mikolov et al., 2013b], namely sub-sampling the frequent terms (using a threshold of 10^{-5} as recommended for large datasets) and smoothing the negative distribution by using probabilities $\{f_i^{3/4}/Z\}$ where f_i is the frequency of word i for sampling negative contexts $\{c'_i\}$.

We then evaluate our embeddings on the following datasets: Simlex [Hill et al., 2015], WordSim [Finkelstein et al., 2002], MEN [Bruni et al., 2014], MC [Miller and Charles, 1991], RG [Rubenstein and Goodenough, 1965], YP [Yang and Powers, 2005], MTurk [Radinsky et al., 2011] [Halawi et al., 2012], RW [thang Luong et al., 2013], using the context embeddings and the Wasserstein-Bures cosine as a similarity measure.

Hypernymy We train our embeddings on the transitive closure of the WORDNET dataset [Miller, 1995] which features 743,241 hypernymy relations on 82,115 distinct nouns. For disambiguation, note that if (u, v) is a hypernymy relation with $u \neq v$, (v, u) is in general *not* a positive relation, but (u, u) is as a noun is always its own hypernym.

We perform our optimization using SGD with batches of 1000 relations, a learning rate 0.02 for dimensions 3 and 4 and 0.01 for higher dimensions, sample 50 negative examples per positive relation, use 6 square root iterations and add $0.01\mathbf{I}$ to the covariances. Contrary to the skipgram experiment, we use a single set of embeddings and use the Wasserstein-Bures dot product as a similarity measure.

Wasserstein-Bures Cosine

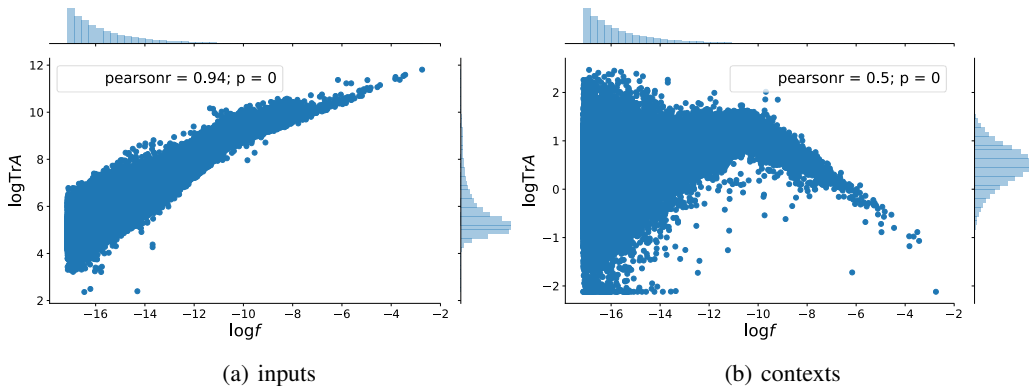


Figure 7: log-log plot of the traces of the embeddings’ covariances vs. word frequency: the sizes of the input embeddings follow a power law, whereas context embeddings give less importance to very frequent words and emphasize on medium frequency words.

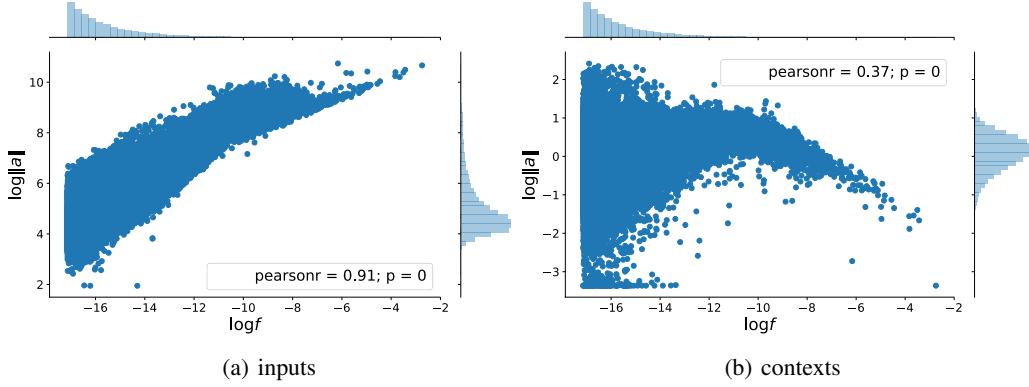


Figure 8: log-log plot of the norms of the embeddings’ means vs. word frequency: the sizes of the input embeddings follow a power law, whereas context embeddings give less importance to very frequent words and emphasize on medium frequency words.

As discussed in section 4, a natural choice of similarity measure would be the Wasserstein-Bures cosine, obtained by normalizing the Wasserstein-Bures dot product with the means’ norms and covariances’ root traces jointly:

$$\text{cos}_{\mathfrak{B}}[\rho_{\mathbf{a},\mathbf{A}}, \rho_{\mathbf{b},\mathbf{B}}] := \frac{\langle \mathbf{a}, \mathbf{b} \rangle + \text{Tr}[\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}}]^{\frac{1}{2}}}{(\|\mathbf{a}\|^2 + \text{Tr} \mathbf{A})^{\frac{1}{2}} (\|\mathbf{b}\|^2 + \text{Tr} \mathbf{B})^{\frac{1}{2}}}$$

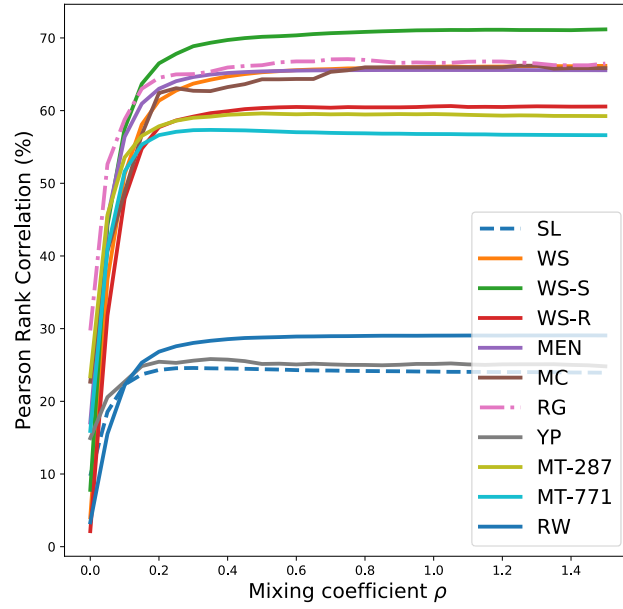
However, we have found that in some applications (and notably in our skipgram experiments) such a joint normalization can result in either the means or the covariances to have a negligible contribution if the scales of the parameters differ too much. To circumvent this problem, we introduce another similarity measure, which is a mixture of two cosine terms:

$$\mathfrak{S}_{\mathfrak{B}}[\rho_{\mathbf{a},\mathbf{A}}, \rho_{\mathbf{b},\mathbf{B}}] := \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\| \|\mathbf{b}\|} + \frac{\text{Tr}[\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}}]^{\frac{1}{2}}}{\sqrt{\text{Tr} \mathbf{A} \text{Tr} \mathbf{B}}}$$

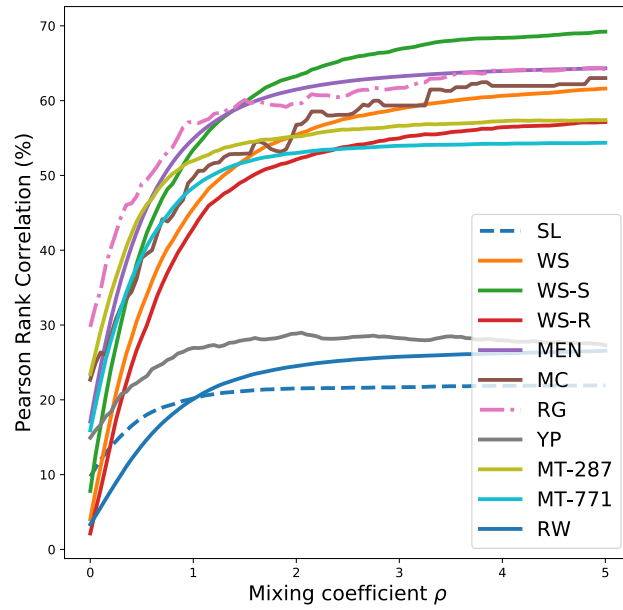
This latter similarity measure allows to gather information from the means and the covariances independently. Note that while the term corresponding to the covariances is obtained in a cosine-like normalization, it takes values between 0 and 1 as it only involve traces of PSD matrices, whereas the means term is a regular Euclidean cosine and therefore takes values between -1 and 1. We compare the behaviors of these two measures on the word similarity evaluation task by introducing a mixing coefficient ρ , and defining

$$\begin{aligned} \text{cos}_{\mathfrak{B}}[\rho_{\mathbf{a},\mathbf{A}}, \rho_{\mathbf{b},\mathbf{B}}; \rho] &:= \frac{\langle \mathbf{a}, \mathbf{b} \rangle + \rho \text{Tr}[\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}}]^{\frac{1}{2}}}{(\|\mathbf{a}\|^2 + \rho \text{Tr} \mathbf{A})^{\frac{1}{2}} (\|\mathbf{b}\|^2 + \rho \text{Tr} \mathbf{B})^{\frac{1}{2}}} \\ \mathfrak{S}_{\mathfrak{B}}[\rho_{\mathbf{a},\mathbf{A}}, \rho_{\mathbf{b},\mathbf{B}}; \rho] &:= \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\| \|\mathbf{b}\|} + \rho \frac{\text{Tr}[\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}}]^{\frac{1}{2}}}{\sqrt{\text{Tr} \mathbf{A} \text{Tr} \mathbf{B}}} \end{aligned}$$

As can be seen from figure 9, the Wasserstein-Bures cosine is less well behaved and makes it difficult to find an optimal mixing value. On the other hand, the mixture of cosines similarity measure varies more smoothly and seems to reach a performance maximum around $\rho = 1$, and achieves better performance than the Wasserstein-Bures cosine on most datasets.



(a) \mathfrak{S}_3



(b) \cos_3

Figure 9: Pearson rank correlation scores on similarity benchmarks as a function of the mixing coefficient: \mathfrak{S}_3 smoothly attains a maximum in performance around $\rho = 1$, whereas \cos_3 has a not so smooth behavior.