A Study of truncated Optimal Topic Transport

Nathan Godey

Abstract

Defining metrics and distances on the space of text documents is a difficult task. One reason for the complexity of this task is the huge dimension of the vocabulary space, namely the important number of different words in a corpus of documents.
A topic model will often be useful to capture recurring topics over a corpus of documents, and will then allow to interpret a document as a distribution over these topics. This leads one to construct *meta-distances*, based on optimal transport between documents as distributions of topics, relying on a precomputed distance between topics, first introduced as Hierarchical Optimal Topic Transport [1]. What we want to address in this report is a fine study of this kind of meta-distances, on different types of underlying distributions. We also want to evaluate the effect of the truncation proposed in the HOTT article [1], both in terms of computation time reduction and of approximation error.

Keywords: optimal transport, document embedding, topic model

1 Introduction

This report is is an attempt to study objects defined as *meta-distances* based on Optimal Transport, such as the HOTT distance. In order to study these distances, we will first define a general framework, which will be useful to cover theoretical explanations; then we will describe the HOTT distance and its more specific setting; finally, we will set up a practical *meta-distance*, which will be more prone to experimentations, without drifting too far away from the HOTT distance.

1.1 General framework: *meta-distances*

Let's consider M, a mixture of T probability distributions over $\mathbb{R}^W$, defined by a set of parameters $(\theta_i, \phi_i)_{i\in[1,T]}$, where θ_i represents the weight of each component $m(\phi_i)$ of the mixture M.

We can define an *inter-component* or *inter-topic* distance as the pairwise k-Wassertein distance W_k between components of M :

$$\forall i, j \in [1, T], d(m(\phi_i), m(\phi_j)) = W_k(m(\phi_i), m(\phi_j))$$

Let's now consider sets $(d_i)_{i\in[1,D]}$ of realizations of the mixture M, of variable sizes, which we are going to call *documents*, to match the HOTT article. More precisely, we can define a document as some $d_i = (w_j)_{j\in[1,|d_i|]}$, where the (w_j) come from a set of realizations $(w_j)_{j\in[1,V]} \sim M$. A straight-forward way of using Optimal Transport is to compute the normalized histograms

$(H(d_i))_{i\in[1,D]}$ of the realizations in the documents, and to compute the k-Wasserstein distances between these histograms (in $\mathbb{R}^V$).
Nevertheless, using our knowledge of the underlying mixture M, we can compute the *a posteriori* probabilities $P(w_j \sim m(\phi_i)|w_j)$ for each component i and for each realization in the document w_j (w stands for word here). Averaging these a posteriori probability distributions over every realization w_j will provide a representation of the document d_i as a distribution over the T components of the mixture M.
However, some topic models provide different ways to get the representation of the documents d_i as distributions over components (or topics), as we will see later on. Now, the documents can be represented as distributions $(\bar{d}_i)_{i\in[1,D]}$ over the T components of M. Our *meta-distance* is going to be a computation of the optimal transport between the representations $(\bar{d}_i)$, based on the underlying inter-component distance that we have discussed earlier.

We define the Component Mover's Distance Matrix C^t as :

$$\forall i, j \in [1, T], C^t_{ij} = W_k(m(\phi_i), m(\phi_j))$$

Thus, the meta-distance $\mathcal{D}$ between two documents d_i and d_j can be defined using the Kantorovitch formulation of Optimal Transport:

$$\mathcal{D}(d_i, d_j) := \min_{P\in\mathbb{R}^{T\times T}_{\geq 0}} \{\langle P, C^t\rangle; P\mathbf{1} = \bar{d}_i, P^T\mathbf{1} = \bar{d}_j\}$$

1.2 An example of meta-distance: Hierarchical Optimal Topic Transport (HOTT)

1.2.1 Prior to HOTT : WMD and Relaxed WMD

When it comes to computing the distance between text documents, one solution is to use the *Word Mover's Distance* or WMD [2]. Basically, WMD solves the Optimal Transport problem between the documents represented as Bag-of-Words, where the underlying distance between words is computed using trained embedding systems (word2vec [3] or GloVe [4] for instance). The WMD problem can be written as :

$$WMD(d_i, d_j) := \min_{P\in\mathbb{R}^{W\times W}_{\geq 0}} \{\langle P, C^v\rangle; P\mathbf{1} = H(d_i), P^T\mathbf{1} = H(d_j)\}$$

where $H(d_i)$ and $H(d_j)$ are normalized bag-of-words (nBOW) representations and C^v stores the pairwise distances between the embeddings of the words in $\mathbb{R}^W$.

However, when the corpora contain a lot of documents, with a lot of different words, WMD faces a complexity issue. Actually, the average time complexity of the basic WMD scales with $O(V^3 log(V))$ where V is the vocabulary size. This issue can be partially solved at the cost of approximations and relaxing constraints. This is what is done with Relaxed WMD (or RWMD).
The RWMD problem is defined as :

$$RWMD(d_i, d_j) := max(\min_{P\in\mathbb{R}^{V_1\times V_2}_{\geq 0}} \{\langle P, C^v\rangle; P\mathbf{1} = H(d_i)\}, \min_{P\in\mathbb{R}^{V_1\times V_2}_{\geq 0}} \{\langle P, C^v\rangle; P^T\mathbf{1} = H(d_j)\})$$

which is the greatest of the WMD problems where one of the constraints is relaxed. Here, it is important that the vocabulary space is constrained to the vocabulary of each document, which means $C^V \in \mathbb{R}^{V_1 \times V_2}$. The authors of the HOTT article show that when the supports of the two documents are close (i.e. they share a lot of words), the RWMD distance plummets, because P costlessly maps each word to itself, regardless of the word distribution.

1.2.2 HOTT

What is proposed with the HOTT distance [1] is to compute a meta-distance between documents, based on a topic model. Thus, the complexity cost is divided in two problems :

- We need to compute the inter-topic distance, seeing the topics as Bags of Words, which is exactly the same as computing the WMD distance on a small corpus.
- We can then compute the meta-distance much more efficiently, because the Component (or Topic) Mover's Distance Matrix C^t is computed once and for all, and its dimension can be controlled as long as we control the number of topics T.

More precisely, the authors use a Latent Dirichlet Allocation [5] model to infer topics from the corpus.

Latent Dirichlet Allocation (or LDA) is a generative probabilistic model of a corpus of text documents. It assumes, for each document d as a bag-of-words vector in a corpus, that it has been generated by :

1. Picking a number of words $N \sim Poisson(\xi)$
2. Picking a "topic combination" $\alpha \sim Dir(\theta)$. This corresponds to our document embedding $\bar{d}$.
3. For each of the N words :
   (a) Picking a topic $z_n \sim Multinomial(\alpha)$
   (b) Picking a word w_n according to the law $p_\beta(w|z_n)$, where β is defined as : $\beta_{i,j} = P(w_i|z_j)$

In the case of the HOTT distance, one fits the LDA topic model to the given corpus D and retrieves a set of topics $(t_i)_{i \in [1,T]} \in \Delta^V$ that are distributions over words, as well as representations of the documents $(\bar{d}_j)_{j \in [1,D]} \in \Delta^T$ as distributions over topics.
We can then define the Topic Mover's Distance Matrix as

$$\forall i, j \in [1,T]^2, C^t_{i,j} := W_1(t_i, t_j) = WMD(t_i, t_j)$$

Finally, we can compute HOTT meta-distance by solving the Optimal Transport problem defined by :

$$\mathcal{D}(d_i, d_j) := \underset{P \in \mathbb{R}^{T \times T}_{\geq 0}}{min} \{\langle P, C^t \rangle ; P\mathbf{1} = \bar{d}_i, P^T \mathbf{1} = \bar{d}_j\}$$

What reduces greatly the complexity of the WMD computation here is the fact that $T \ll V$, i.e. we expect to find much less topics than there are unique words in the corpus.

1.3 A dummy meta-distance

HOTT distance relies on real data, namely text documents, which enforces a costly computation when fitting a topic model, and very poor control over the distributions and the dimensionality of the considered spaces (vocabulary and topics).
In order to gain control over these parameters, we are going to pick a specific mixture M, namely a Gaussian Mixture of parameters $(\theta_i, \mu_i, \Sigma_i)_{i\in[1,T]}$ in the space $\mathbb{R}^W$.
Since the (μ_i) will model distributions over the vocabulary space (as an analogy to the (t_i)), we ensure that $\Sigma_{k=1}^V \mu_{i,k} = 1$ by normalization, and that $\mu_i \geq 0$ for every i. We can now pick T, V, and D, and also choose the structure of our mixture M by picking its parameters.
Another advantage of this approach is that we have a closed form for the W_2 distance between two multivariate Gaussians.

Property 1.1 *Let's consider $N_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $N_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$. One has:*

$$W_2(N_1, N_2)^2 = \|\mu_1 - \mu_2\|_2^2 + \mathrm{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2})$$

This allows us to introduce the Topic Mover's Distance Matrix as:

$$\forall i, j \in [1, T]^2, C_{i,j}^{\mathcal{N}} = \sqrt{\|\mu_i - \mu_j\|_2^2 + \mathrm{Tr}(\Sigma_i + \Sigma_j - 2(\Sigma_i^{1/2}\Sigma_j\Sigma_i^{1/2})^{1/2})}$$

The last thing we need to do in order to create our dummy meta-distance is to find a way to generate the "documents" $(d_i)_{i\in[1,D]}$ and their representations $(\bar{d}_i)$ as distributions in Δ^T. We can use the vector made of the weights of the mixture $\boldsymbol{\theta} = (\theta_i)_{i\in[1,T]}$ to that purpose, by taking inspiration from the LDA :

$$\forall i \in [1, D], \bar{d}_i \sim Dir(\frac{1}{\sigma}\theta)$$

We use a Dirichlet distribution for the documents since it draws a vector in the simplex Δ^T, and ensures that $E(\bar{d}_i) = \boldsymbol{\theta}$, which makes sense regarding the analogy with text documents (the average representation of the documents reflects the repartition θ of the topics in the corpus).
The σ parameter allows us to control the magnitude of the covariance of $\bar{d}_i$:

$$Cov(\bar{d_i}^k, \bar{d_i}^l) = \frac{\sigma\delta_{kl}\theta_k - \sigma^2\theta_k\theta_l}{1 + \frac{1}{\sigma}} \sim \sigma\delta_{kl}\theta_k - \sigma^2\theta_k\theta_l$$

We can now detail the process of generating dummy data for our meta-distance given T, V, D, and σ:

1. Draw T sets of parameters $(\theta_i, \mu_i, \Sigma_i)$ for the mixture, where $\theta_i \in \mathbb{R}^T$ and $\mu_i \in \mathbb{R}_{\geq 0}^V$ are normalized.
2. Draw D documents as distributions over the components of the mixture $(\bar{d}_i)$ having $\bar{d}_i \sim Dir(\frac{1}{\sigma}\theta)$

Then, we can estimate the embedding of a document in the space $\mathbb{R}^V$, as an analogy to its bag-of-words representation, using $d_i := \Sigma_{k=1}^T \bar{d_i}^k \mu_i^k$.
Our dummy meta-distance thus solves the following problem :

$$\mathcal{D}(d_i, d_j) := \min_{P\in\mathbb{R}_{\geq 0}^{T\times T}} \{\langle P, C^{\mathcal{N}}\rangle; P\mathbf{1} = \bar{d}_i, P^T\mathbf{1} = \bar{d}_j\}$$

1.4 Contributions

This report aims at adding personal ideas and conduct experiments that are not presented in the studied litterature. First, as seen above, we tried to introduce a general framework for what is introduced as meta-distances in [1]. To study this kind of distances, we introduced a dummy meta-distance on the GMM model. We adapt the demonstrations in [1] to this general framework, to show that meta-distances are actually distances, and to write an inequalty between WMD-like distances and meta-distances. We also try to find a satisfying bound on the approximation error due to the truncation of HOTT. We conduct experiments on the behaviour of meta-distances to check the hypothesis presented in [1]. We show empirically that meta-distances indeed seem to provide better semantical representations when looking at the documents in the topic space. We investigate the argument undermining RWMD, and check the pertinence of the choice of the threshold for truncated HOTT via experimentation.

2 Studying meta-distances

2.1 Theoretical study

In this section, we aim at exploring some properties of meta-distances in their broad framework.

2.1.1 Is a meta-distance a distance?

Computing a meta-distance amounts to solving a discrete Kantorovitch problem on the topic space $\mathbb{R}^T$. This leads us to think that there is a way to formulate the meta-distance problem into a Wasserstein distance computation.
In the article that introduces HOTT [1], the authors achieve to formalize the meta-distance problem into a Wasserstein distance computation, using a representation based on Dirac distributions:

$$HOTT(d_i, d_j) = W_1(\sum_{k=1}^{T} \bar{d_i^k} \delta_{t_k}, \sum_{k=1}^{T} \bar{d_j^k} \delta_{t_k})$$

Here, the underlying distance is :

$$d(\delta_{t_i}, \delta_{t_j}) = WMD(t_i, t_j)$$

In our general framework, we can introduce Dirac distributions (δ_{ϕ_k}), where (ϕ_k) are the parameters of each component $m(\phi_k)$ of the mixture. Our underlying k-Wasserstein distance will be:

$$d(\phi_i, \phi_j) = W_k(m(\phi_i), m(\phi_j))$$

We can now explicit our meta-distance $\mathcal{D}$ as a 1-Wasserstein distance:

$$\mathcal{D}(d_i, d_j) = W_1(\sum_{k=1}^{T} \bar{d_i^k} \delta_{\phi_k}, \sum_{k=1}^{T} \bar{d_j^k} \delta_{\phi_k})$$

This means that a meta-distance is in fact a distance. This yields the idea of extending the idea of meta-distances by computing the k-Wasserstein distance, which can be more relevant with respect

to the underlying distance.
In the case of our dummy distance for instance, where the underlying distance is a 2-Wasserstein on the space of multivariate Gaussian distributions, we could use :

$$\mathcal{D}_2(d_i, d_j) = W_2(\sum_{k=1}^{T} \bar{d}_i^k \delta_{\mu_k,\Sigma_k}, \sum_{k=1}^{T} \bar{d}_j^k \delta_{\mu_k,\Sigma_k})$$

2.1.2 Relation between a meta-distance on distributions and the histogram distance

Let's pick a specific case where the expectations of the components of the mixture $(E(m(\phi_i)))$ are themselves discrete distributions of $\mathbb{R}^T$, i.e. we have:

$$\forall i \in [1, T], \sum_{k=1}^{T} E(m(\phi_i))_k = 1 \text{ and } E(m(\phi_i)) \geq 0$$

This is the case in both the HOTT article and our dummy distance. We have an approximation of d_i which is:

$$d_i \simeq \sum_{k=1}^{T} \bar{d}_i^k E(m(\phi_k))$$

In the HOTT article, where the mixture is a topic model made of the topics (t_k) this comes as:

$$d_i \simeq \sum_{k=1}^{T} \bar{d}_i^k t_k$$

Now, let's consider the histogram distance we've discussed in the introduction using W_1, which is basically a generalization of WMD :

$$d(H(d_i), H(d_j)) = W_1(H(d_i), H(d_j))$$

Using the triangular inequalty:

$$\begin{aligned} d(H(d_i), H(d_j)) \leq & W_1(H(d_i), \sum_{k=1}^{T} \bar{d}_i^k E(m(\phi_k))) \\ & + W_1(\sum_{k=1}^{T} \bar{d}_i^k E(m(\phi_k)), \sum_{k=1}^{T} \bar{d}_j^k E(m(\phi_k))) \\ & + W_1(\sum_{k=1}^{T} \bar{d}_i^k E(m(\phi_k)), H(d_j)) \end{aligned} \tag{1}$$

In order to go further, we need to invoke two lemmas.

Lemma 2.1 *Let p and q be two probability distributions in $\mathbb{R}^N$. Then we have:*

$$|E(p) - E(q)| \leq W_1(p, q)$$

Lemma 2.2 (Derived from Talagrand inequalty, see Otto&Villani [6]) *Let p and q be distributions on a finite-diameter metric space X. Then:*

$$W_1(p,q) \leq diam(X)\sqrt{\frac{1}{2}KL(p||q)}$$

First, lemma 2.2 lets us bound two terms of the inequalty:

- $W_1(H(d_i), \sum_{k=1}^{T} \bar{d_i^k} E(m(\phi_k))) \leq C\sqrt{KL(H(d_i)|| \sum_{k=1}^{T} \bar{d_i^k} E(m(\phi_k)))}$
- $W_1(\sum_{k=1}^{T} \bar{d_j^k} E(m(\phi_k)), H(d_j)) \leq C\sqrt{KL(\sum_{k=1}^{T} \bar{d_j^k} E(m(\phi_k))||H(d_j))}$

The middle term can be shown to be bounded by our meta-distance, as we see in the following property.

Property 2.1 *We have:*

$$W_1(\sum_{k=1}^{T} \bar{d_i^k} E(m(\phi_k)), \sum_{k=1}^{T} \bar{d_j^k} E(m(\phi_k))) \leq W_1(\sum_{k=1}^{T} \bar{d_i^k} \delta_{\phi_k}, \sum_{k=1}^{T} \bar{d_j^k} \delta_{\phi_k}) = \mathcal{D}(d_i, d_j)$$

Proof: Let's write down the two 1-Wasserstein distances above as Kantorovitch problems. For the left-hand term, we can define the cost matrix C^{left} as:

$$\forall i,j \in [1,T]^2, C_{i,j}^{left} = |E(m(\phi_i)) - E(m(\phi_j))|$$

For the right-hand term, we have by definition of the underlying distance of $\mathcal{D}$:

$$\forall i,j \in [1,T]^2, C_{i,j}^{right} = W_1(m(\phi_i), m(\phi_j))$$

Lemma 2.1 then allows us to show that :

$$\forall i,j \in [1,T]^2, C_{i,j}^{left} \leq C_{i,j}^{right}$$

Therefore, comparing the Kantorovitch formulations becomes easy, because any eligible transport plan P on the right-hand side will be **eligible and less costly** on the left-hand side:

$$\underset{P\in\mathbb{R}_{\geq 0}^{T\times T}}{min} \{\langle P, C^{left}\rangle; P\mathbf{1} = \bar{d_i}, P^T\mathbf{1} = \bar{d_j}\} \leq \underset{P\in\mathbb{R}_{\geq 0}^{T\times T}}{min} \{\langle P, C^{right}\rangle; P\mathbf{1} = \bar{d_i}, P^T\mathbf{1} = \bar{d_j}\}$$

Switching back to 1-Wasserstein formulations allows us to finish the proof.

Combining what we have done above yields the following property:

Property 2.2 *There exists a constant C relying on the finite diameter of the support of the distributions only, such that:*

$$W_1(H(d_i), H(d_j)) \leq \mathcal{D}(d_i, d_j) + C(\sqrt{KL(H(d_i)|| \sum_{k=1}^{T} \bar{d_i^k} E(m(\phi_k)))} + \sqrt{KL(\sum_{k=1}^{T} \bar{d_j^k} E(m(\phi_k))||H(d_j)))}$$

This property is useful when dealing with the HOTT distance, because the LDA topic model minimizes the Kullback-Leibler divergences on the right-hand side, and does even more so as T increases.

2.1.3 A bound on the approximation error due to the truncated HOTT

In the HOTT article [1], the authors introduce the idea of truncated Optimal Topic Transport. The idea is to set a threshold ν and apply the corresponding (zero out coordinates below ν) to the documents' representations $(\bar{d_i})$, followed by a normalization. We note $(\tilde{d_{\nu,i}})$ the results of this thresholding operation.
The advantage of this thresholding operation is that the resulting distributions become sparser, which reduces computation time when solving the Optimal Transport problem.

Our goal in this section is to provide a bound to the approximation error that derives from this thresholding when the $(\bar{d_i})$ are distributed according to a Dirichlet distribution. The triangular inequalty gives us:

$$\Delta_{thresh} = |W_1(\bar{d_i}, \bar{d_j}) - W_1(\tilde{d_{\nu,i}}, \tilde{d_{\nu,j}})| \leq W_1(\bar{d_i}, \tilde{d_{\nu,i}}) + W_1(\bar{d_j}, \tilde{d_{\nu,j}})$$

Hence, we want to bound $W_1(\bar{d_i}, \tilde{d_{\nu,i}})$ for every i.
Let's define $n(\bar{d_i}, \nu)$ the number of coordinates of a given $\bar{d_i}$ that are below the threshold ν. Since $\bar{d_i} \sim Dir(\theta)$, we have $\bar{d_i^k} \sim Beta(\theta_k, \theta_0 - \theta_k)$ with $\langle \theta, \mathbf{1} \rangle = \theta_0$.

The cumulative distribution function of the $Beta(\alpha, \beta)$ distribution is the regularized incomplete beta function $I(x; \alpha, \beta)$. Thus:

$$\forall k \in [1, T], P(\bar{d_i^k} \leq \nu) = I(\nu; \theta_k, \theta_0 - \theta_k)$$

This expression yields:

$$E(n(\bar{d_i}, \nu)) = \sum_{k=1}^{T} P(\bar{d_i^k} \leq \nu) = \sum_{k=1}^{T} I(\nu; \theta_k, \theta_0 - \theta_k)$$

Now, let's focus on $W_1(\bar{d_i}, \tilde{d_{\nu,i}})$. Here, we aim at transporting the mass of the coordinates that are smaller than ν to the ones that are greater. Hence, we need to transport $n(\bar{d_i}, \nu)$ masses that weigh at most ν at a cost of at most $||C||_\infty$, where C is the inter-topic cost matrix. If we formalize this, we get:

$$W_1(\bar{d_i}, \tilde{d_{\nu,i}}) \leq \nu n(\bar{d_i}, \nu) ||C||_\infty$$

If we use this result in our previous inequalty, we obtain the following bound:

$$\Delta_{thresh} \leq W_1(\bar{d_i}, \tilde{d_{\nu,i}}) + W_1(\bar{d_j}, \tilde{d_{\nu,j}}) \leq \nu ||C||_\infty (n(\bar{d_i}, \nu) + n(\bar{d_j}, \nu))$$

Using the previous result on $E(n(\bar{d_i}, \nu))$, we can provide a bound to $E(\Delta_{thresh})$:

$$E(\Delta_{thresh}) \leq 2\nu ||C||_\infty \sum_{k=1}^{T} I(\nu; \theta_k, \theta_0 - \theta_k)$$

This result can be used directly on the HOTT distance, provided we retrieve the θ fitted by the LDA topic model. Moreover, in the case of our dummy distance, where the variance of $\bar{d_i}$ increases with

σ, this result can be rewritten:

$$E(\Delta_{thresh}) \leq 2\nu ||C||_\infty \sum_{k=1}^{T} I(\nu; \frac{\theta_k}{\sigma}, \frac{1-\theta_k}{\sigma})$$

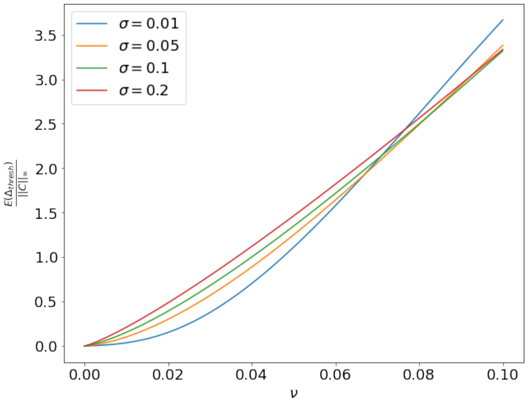


FIGURE 1: Bound of $\frac{E(\Delta_{thresh})}{||C||_\infty}$ for various values of σ ($T = 20$)

2.2 Meta-distance on the Gaussian Mixture Model

2.2.1 General study

In this section, we will study the behaviour of our dummy meta-distance, built on the Gaussian Mixture Model. In a basic setup, we can observe how the different distributions behave, as we see in the following figures. We used a Monte-Carlo method with 1000 realizations for each document d_i to compute their nBoW-like representations.

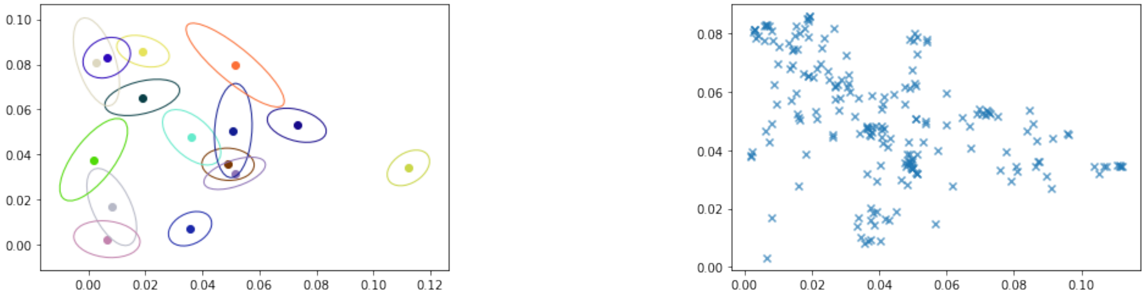


FIGURE 2: Distributions in the underlying space, projected on the first two dimensions. On the left, we see the Gaussian Mixture's components' means and ellipses to model their covariances. On the right, we observe 200 documents as averages of their realizations, which should more or less belong in the convex hull of the components' means. *Parameters:* $T = 15$, $V = 20$, $\sigma = 2$, $||\Sigma_i|| \sim 10^{-5}$

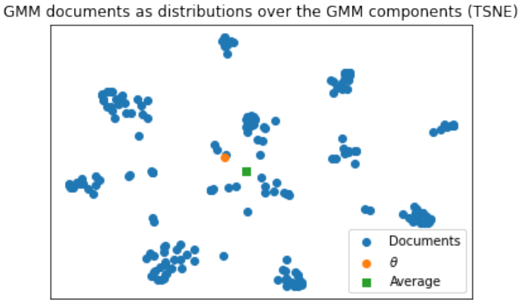


FIGURE 3: t-SNE of the documents as distributions over the components of the GMM of weight parameter θ

Now that we have found a way to represent the different spaces, the topics, and the documents, we can try to display the meta-distance. Using the closed form for W_2 between Gaussian multivariate distributions, we can first compute the Component Mover's Cost Matrix $C^{\mathcal{N}}$.

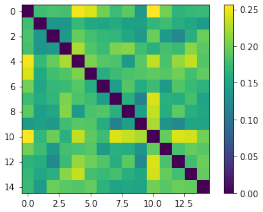
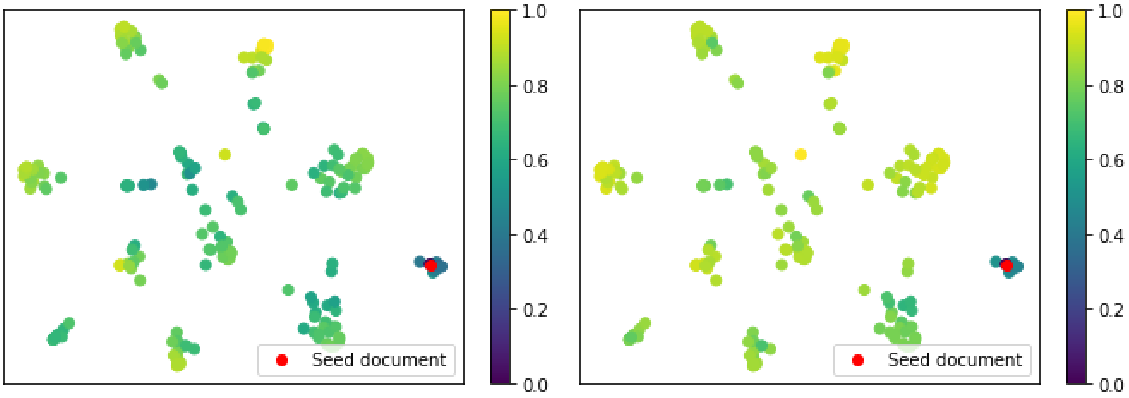


FIGURE 4: Component Mover's Cost Matrix $C^{\mathcal{N}}$

In order to visualize the meta-distance, we are going to randomly pick a seed document, and we are going to compute the meta-distance between that seed document and every other document.



(A) WMD-like distance

(B) Meta-distance

FIGURE 5: Distances to a seed document on tSNE projections in the topic space. *Parameters: as above*

We can notice that the WMD-like distance, on the left, provides small values when comparing documents that belong in the "center" of the component/topic space with the seed. Moreover, it shows a strong gradient of distance inside the clusters in the component/topic space.

On the other hand, the meta-distances between the seed and the documents belonging in the same cluster of the tSNE seem small compared to every other meta-distance. This shows that the meta-distance might be useful when computing a clustering of the documents. There does not seem to be strong gradients of distance to the seed within the far-away clusters themselves for the meta-distance. Let's try different characteristical set-ups for the meta-distance.

2.2.2 Broad components/topics

When there are not enough components to capture the semantic pattern of the underlying space, it is possible that the component/topic model fits broad components (or vague topics). In our dummy model, it would mean we are trying to look at what happens when the magnitude of the Σ_i parameters of the Gaussian Mixture Models increase. Let's observe our documents and mixture first.

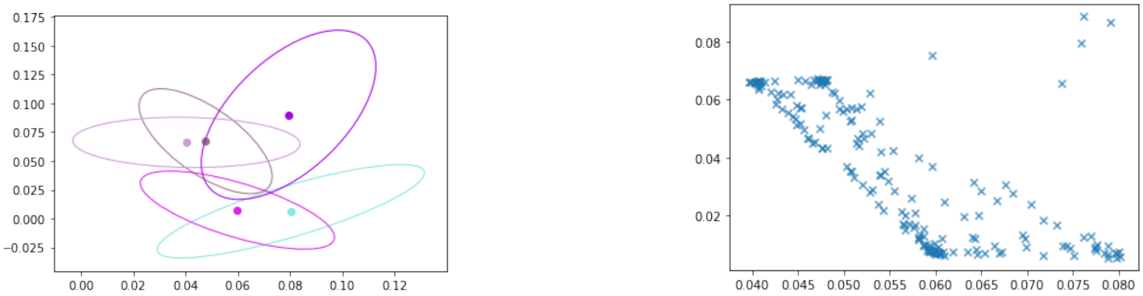


FIGURE 6: Distributions in the underlying space, projected on the first two dimensions. We see here that the components overlap a lot, and that the documents tend to be closer to the means of the components, as σ is big w.r.t. the number of components. *Parameters:* $T = 5$, $V = 20$, $\sigma = 2$, $||\Sigma_i|| \sim 10^{-4}$

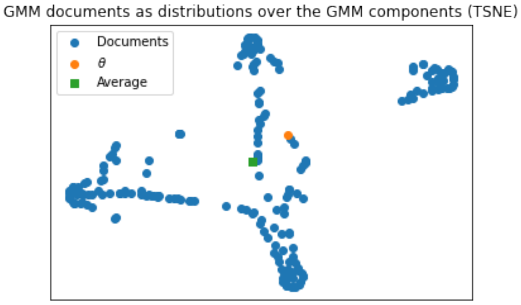


FIGURE 7: t-SNE of the documents as distributions over the components of the GMM of weight parameter θ. We distinguish 4 clusters, which means two components were "merged" together.

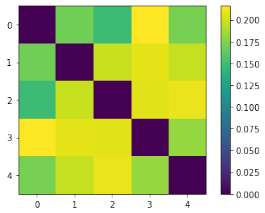
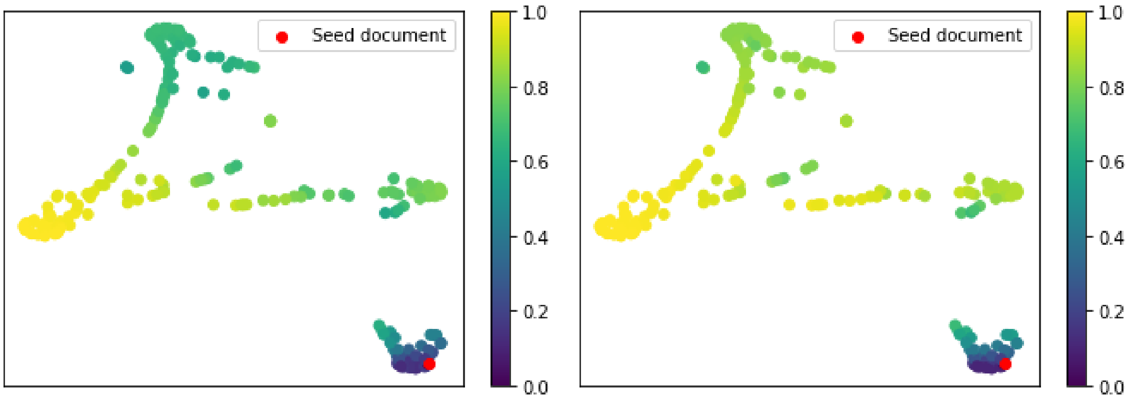


FIGURE 8: Component Mover's Cost Matrix $C^{\mathcal{N}}$



(A) WMD-like distance

(B) Meta-distance

FIGURE 9: Distances to a seed document on tSNE projections in the topic space. *Parameters: as above*

Once again, we see that the meta-distance is better at distinguishing different clusters in the topic space, i.e. two documents from the same cluster are significantly closer to each other than they are to documents from other clusters.

2.2.3 Flawed clustering : marginal components/topics

When a clustering algorithm does not achieve to capture the structure of a point cloud, it can happen that some big predicted clusters catch an important part of the point cloud, while many small clusters try to catch the marginal points. This can be modelled in our dummy model by altering the parameters of the mixture so that specific components are the major ones, and shrinking the sizes of the other ones.

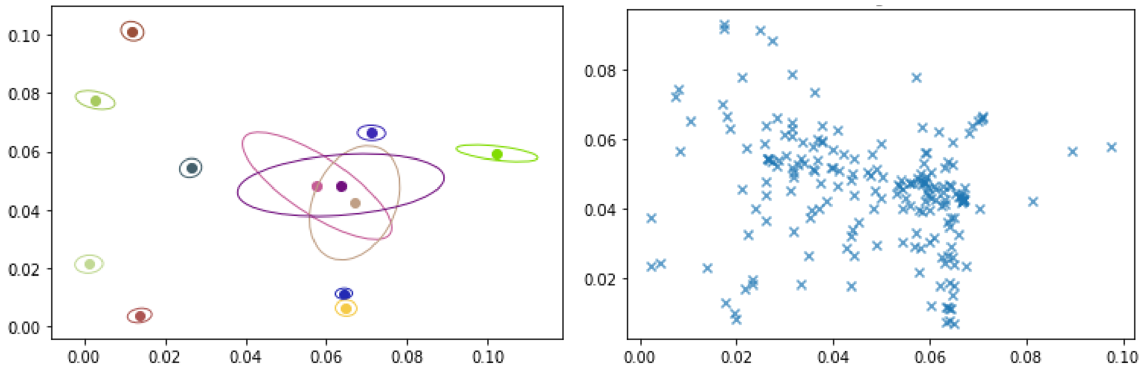


FIGURE 10: Distributions in the underlying space, projected on the first two dimensions. We see three major components on the left, surrounded by small marginal ones. *Parameters:* $T = 12$, $V = 20$, $\sigma = 1$, $||\Sigma_i|| \sim 10^{-5}$

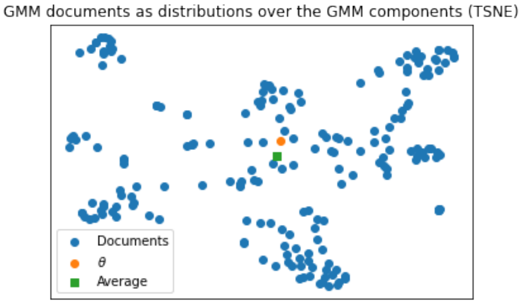
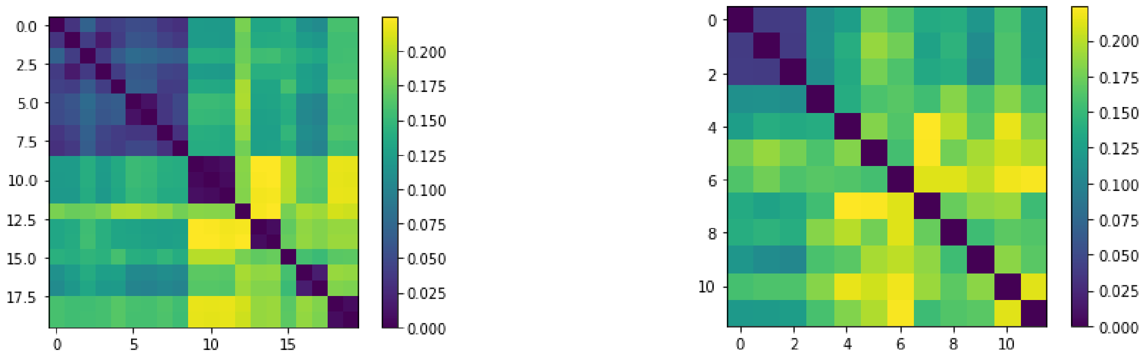


FIGURE 11: t-SNE of the documents as distributions over the component of the GMM

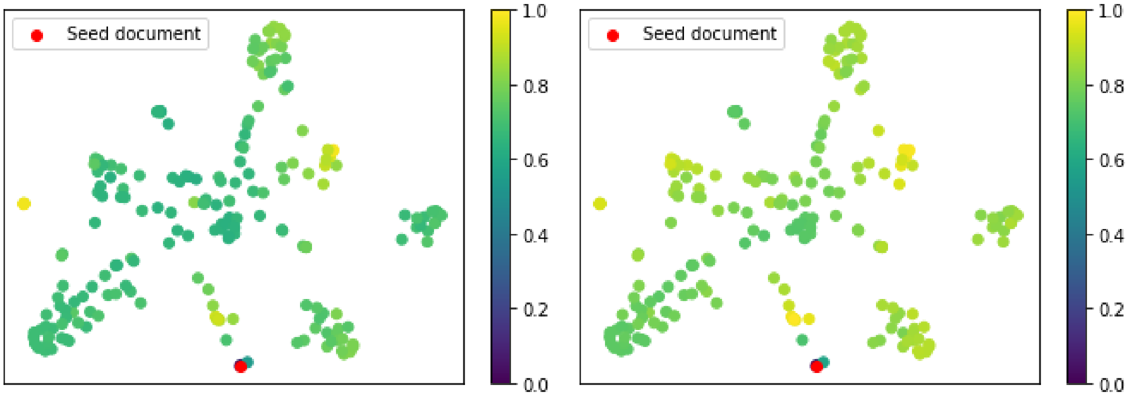


(A) WMD-like distance (inter-"words")

(B) Meta-distance (inter-components)

FIGURE 12: Cost Matrices

On the left-hand side, we notice that some words are very similar to one another, by looking at the squares along the diagonal. On the right-hand side, the Component Mover's Cost Matrix shows very small distance between the three first components (which are the major ones), which are also relatively close to the marginal components. Each marginal component seems fairly far away from the other ones though.

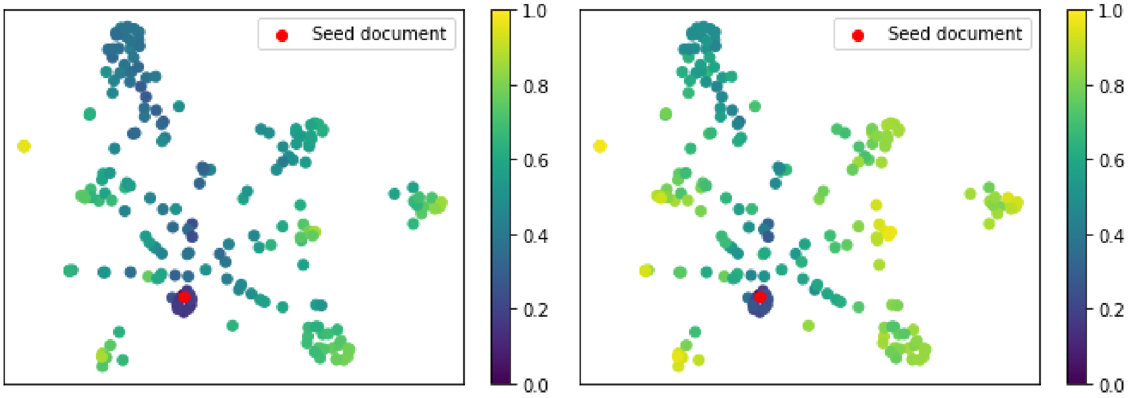


(A) WMD-like distance

(B) Meta-distance

FIGURE 13: Distances to a **marginal** seed document on tSNE projections in the topic space

Once more, we observe that the distance increases slightly faster as we move away from the seed in the case of the meta-distance. How does the meta-distance behave when the document belongs in the major components ?



(A) WMD-like distance

(B) Meta-distance

FIGURE 14: Distances to a seed document **in a major component** on tSNE projections in the topic space (same documents, different tSNE projection)

For the WMD-like distance, we see that the distance to the seed is low for a lot of documents in the center, which are probably documents belonging in the major components too. For the meta-distance, this phenomenon is less important, as we can notice by looking at the cluster at the top of the tSNE projection, which seems more distant than in the case of the WMD-like distance.

Overall, we have seen in this section that in different scenarios of document distributions and for different parameters of the mixture, the meta-distance seems to be more efficient than the WMD-like distance at distinguishing clusters in the topic space.

For that matter, these experiments lead to think that a meta-distance is semantically consistent **at the document scale**, which means that it is able to take advantage of the knowledge of the underlying document distribution (via the mixture / topic model), and to evaluate a distance accordingly to the topics/components we can fit to that distribution. This is a difference when comparing to WMD, which uses the word scale to compare documents, without taking into account the specific document distribution in a corpus.

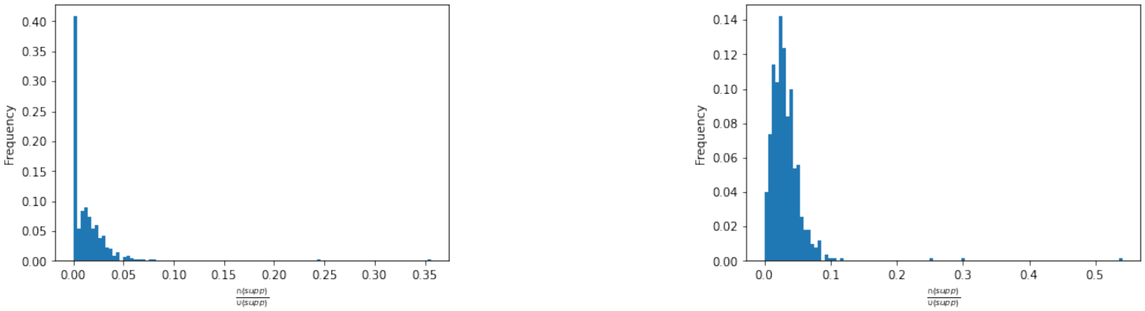
2.3 The HOTT meta-distance

2.3.1 How bad is really RWMD ?

In the HOTT article [1], the authors argue that the Relaxed Word Mover's Distance presented in [2] present flaws when it comes to comparing documents. Their first point is that documents often share a lot of words, i.e. that the support of the normalized Bag-of-Words distributions representing the documents overlap. Assuming this hypothesis on the corpus' structure makes RWMD a bad distance, because relaxing one constraint on the marginals of the transport plan means that P can transport the whole mass of shared words costlessly, regardless of the frequency of these shared words in both documents. In this section, we aim at answering these questions :

- To what extent do natural language documents in a corpus actually share words ? In other terms, are the supports of the nBoW representations actually overlapping ?
- To what extent RWMD is a bad approximation of WMD when the supports of the nBoWs tend to overlap ?

To answer the first question, we are going to use the Enron e-mail dataset, and the KOS blog posts dataset, already cleaned and processed into nBoW representations. For a number of pairs of nBoW representation, we compute the ratio between the intersection and the union of their supports.



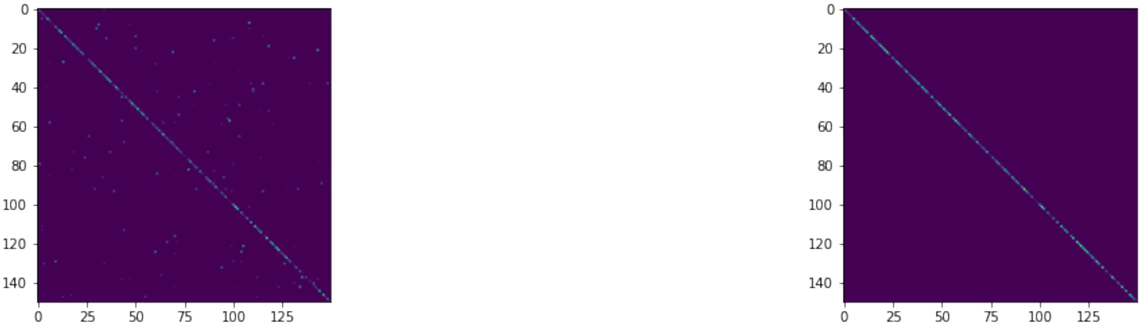
(A) Enron dataset (average: 0.013)

(B) KOS dataset(average: 0.033)

FIGURE 15: Histograms of the intersection rate of 500 sampled nBoW representations

Other corpora might yield different results, but we see here that the assumption that nBoW representations of documents supports overlap is not always true.

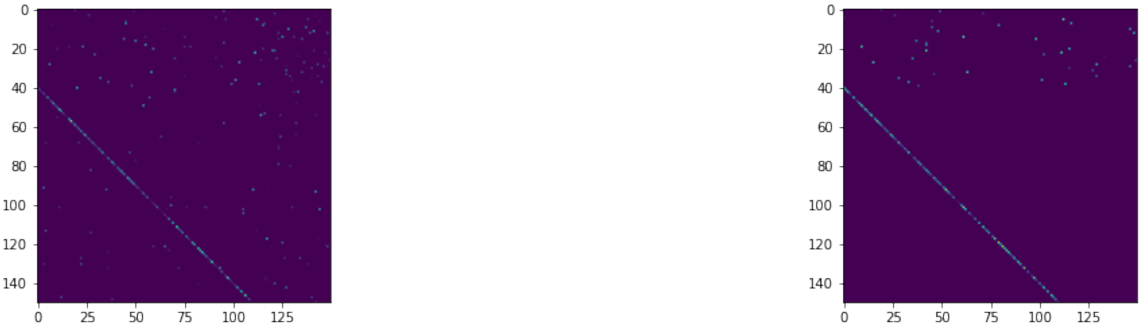
To answer the second question, we generate a dummy problem, where the matrix C is random, and we generate random nBoW representations on given supports. Thus, we can choose the extent to which the nBoW representations overlap. Then, we implement a RWMD solver using linear programming and CVXPY.



(A) WMD

(B) RWMD

FIGURE 16: Optimal transport plans when the supports fully overlap



(A) WMD

(B) RWMD

FIGURE 17: Optimal transport plans when the supports partially overlap

On the figures above, we see how RWMD optimal transport plans are approximations of the ones from WMD. The main difference is that on the support intersections, RWMD is able to transfer the whole mass costlessly, which we clearly see on the "diagonal" shapes, where one of the marginals becomes a mapping from a word to itself. Since RWMD is a maximum between relaxations of WMD, we have:

$$RWMD(d_1, d_2) \leq WMD(d_1, d_2)$$

An interesting thing to study will then be $\frac{RWMD(d_1,d_2)}{WMD(d_1,d_2)}$, with respect to the amount of intersection of the supports of $nBoW(d_1)$ and $nBoW(d_2)$.

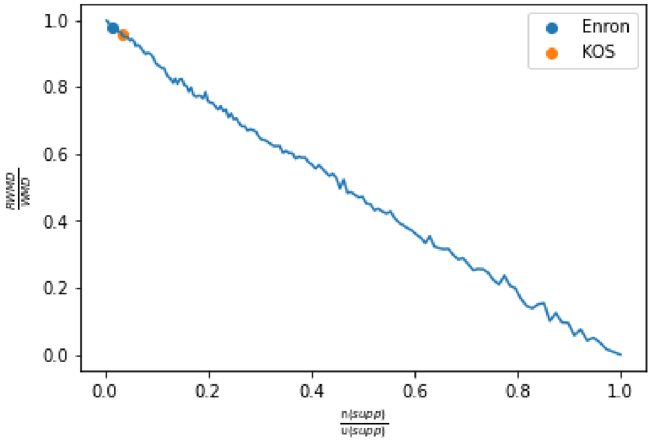


FIGURE 18: Approximation error of RWMD with respect to the intersection rate of the supports of the documents nBoW

As expected, when the supports don't intersect, RWMD and WMD yield the same distance. Also, when the support fully intersect, RWMD becomes null. What is not discussed in [1] is the behaviour when the intersection is partial. Here, we see that the evolution of the relative approximation error between full support intersection and disjoint supports is close to linear. Moreover, when interpolating the average support intersection rate for natural language datasets, we see that RWMD is expected to yield results close to WMD for such corpora. As a conclusion, we can answer our two initial questions:

- It seems that pairs of natural language documents don't actually share that many meaningful words, once cleaned of stop words. Hence, the supports of their nBoW representations don't overlap much (pairwise).
- RWMD approximation quality decays slowly with the intersection rate of the supports of the nBoW representations. Moreover, at support intersection rates we observe in natural language documents, the relative approximation error made by RWMD is supposed to be quite small.

The arguments advanced in [1] to explain why RWMD is a bad distance don't seem to hold considering our experiments.

2.3.2 Effect of truncating

Let's now focus on the HOTT meta-distance itself, on natural language documents. A very important argument in [1] is that truncating the representations of the documents as distributions over topics ($\bar{d}_i$) allows to compute meta-distances faster, with a similar quality when testing on classification tasks. For this section, we will be using and modifying the implementation of HOTT provided by the authors on GitHub [7].

We want to study more precisely the difference between truncated HOTT, and non-truncated HOTT (which is called HOFTT for Hierarchical Optimal Full Topic Transport). In order to do that, we are going to fit an LDA model the CLASSIC dataset, tested in [1]. Then, we want to sample documents as topic distributions, and average the error made by truncated HOTT with respect to HOFTT for different values of the threshold ν, in the range of $\frac{1}{T+1}$, which is the threshold recommended in [1].

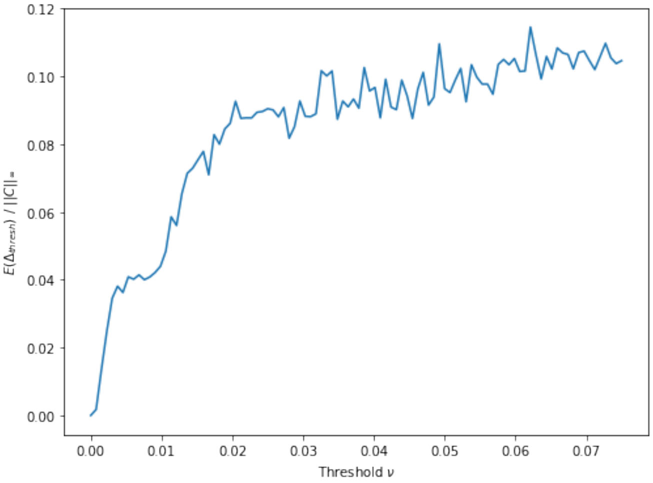


FIGURE 19: Approximation error due to the truncation (as introduced in the Theoretical Study) for various values of the threshold. *Parameters: $T = 40$, $\theta = 0.1 \times \mathbf{1}$, 250 pairs of documents sampled for each threshold)*

We see that for small thresholds, the error increases rapidly. However, once the threshold is greater than $\frac{1}{T+1}$, it seems that the error increases much more slowly. The magnitude of the error w.r.t. $||C||_\infty$ is relatively low (around 0.1). We also see that the bound we introduced earlier is very loose.

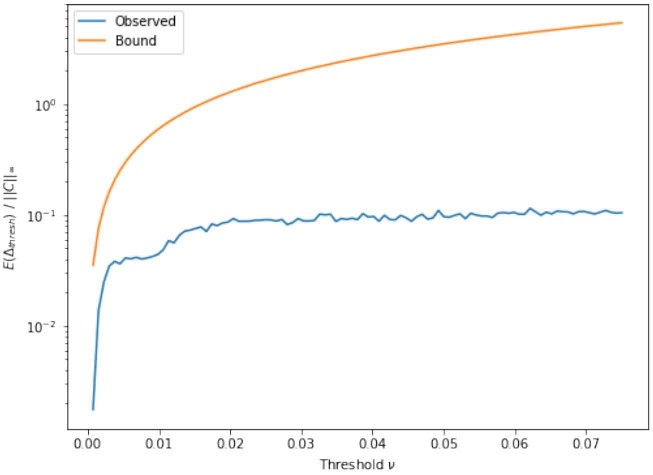


FIGURE 20: Approximation error due to the truncation compared with the theoretical bound (semi-log scale)

Another factor we need to compare is the computation time for both distances, for different threshold levels.

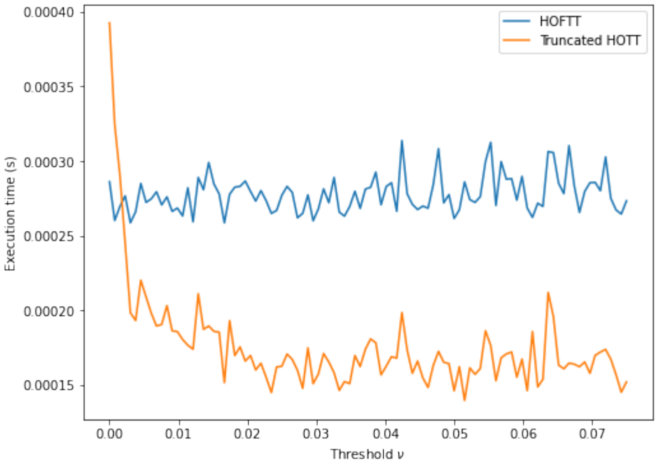


FIGURE 21: Execution time for the HOFTT and truncated HOTT meta-distances

As expected, the truncated HOTT is faster than the HOFTT distance. However, the time advantage stops increasing around $\frac{1}{T+1}$, where truncated HOTT is almost half as fast as the HOFTT.

Overall, we observe that an interesting trade-off is obtained when $\nu \simeq \frac{1}{T+1}$: below that threshold, the execution time increases; above that threshold, the approximation error slightly increases.

3 Conclusion

We've extended the demonstrations of the HOTT article [1] to a more general framework that could be useful in other applications. For instance, in patch-based Image analysis, Zoran and Weiss [8] introduced a Gaussian Mixture Model to capture the patch distribution of real images. If we consider images as sets of realizations of the patch distribution, we clearly see that a meta-distance could be used to compare images based on this work.

We've tried to introduce an upper bound to the approximation error that is done when computed truncated meta-distances. As we've seen though, this bound is rather loose, and doesn't allow to provide a good estimation of the actual approximation error. An idea for future work could be to improve this bound using our knowledge of the constrained space within which our distribution lie, and properties of the optimal transport plan P.

Then, using our dummy model based on a GMM, we showed empirically, in various use cases, that meta-distances tend to differentiate clusters in the topic/component space, rather than in the word/underlying space.

However, it seems that the argument presented in [1] advocating that the RWMD is a bad distance because of overlapping supports of the nBoW representations does not seem to hold. We've empirically seen that natural language corpora are not always containing documents that share lots of words pairwise. This means that the supports of pairs of nBoW representations don't overlap much in general. Moreover, the relaxation introduced by RWMD generates some error, but this error seems to increase slowly with how much the supports intersect. All in all, these experiments showed that this argument is flawed.

Finally, we observed the effect of truncating on the HOTT distance directly. We saw that the provided threshold ($\frac{1}{T+1}$) seems to be close to a good trade-off between computation time improvements and approximation quality.

References

[1] Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, Justin Solomon. *Hierarchical Optimal Transport for Document Representation*, 2019.

[2] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, Kilian Q. Weinberger. *From Word Embeddings To Document Distances*, 2015.

[3] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space*, 2013.

[4] Jeffrey Pennington, Richard Socher, Christopher D. Manning *GloVe: Global Vectors for Word Representation*, 2014.

[5] David M. Blei, Andrew Y. Ng, Michael I. Jordan. *Latent Dirichlet Allocation*, 2003.

[6] Félix Otto, Cédric Villani *Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality*, 2000.

[7] Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, Justin Solomon. *Github repository for the HOTT article*

[8] Daniel Zoran, Yair Weiss *Natural Images, Gaussian Mixtures and Dead Leaves*, 2012.