Report on Hierarchical Optimal Transport for Document Representation

Louise Durand-Janin
ENS Paris-Saclay, Gif-sur-Yvette, France
ldurandjanin@gmail.com

January 2024

Abstract

Being able to compute an efficient and interpretable metric to characterize distance between document is a crucial challenge. While the performances of Deep Learning algorithm to keep track of words semantic have significantly increased, existing metrics fail to combine interpretability and efficiency. The article *Hierarchical Optimal Transport for Document Representation* proposes to leverage the hierarchical structure of corpus of documents as well as word embeddings to adress this problem. While demonstrating comparable performance to past metric, it proves to reduce greatly the computation cost while being human-understandable. It applies the approach of Hierarchical Optimal Transport to document classification. This theory is based on the idea of identifying structures among data to compute the Kantorovitch problem in two separate steps.

1 Introduction

Understanding and measuring the similarities between documents is a fundamental task in Natural Language Processing. It has demonstrated a broad range of application ranging from information retrieval to document classification and summarization of document collections.

Previously introduced distance metrics often face challenges in providing a meaningful semantic interpretation of words within documents or suffer from computational inefficiencies.

1.1 Notations

Let us establish the notations that will be employed throughout this report. Let V denotes the vocabulary and $|V|$ be the length of the vocabulary

- **vocabulary** V : number of unique words in a corpus and has length $|V|$

- **word** : vector $w \in \mathbb{R}^{|V|}$
- **document** : sequence of N words $\mathbf{d} = (w_1, ..., w_N)$
- **corpus** : collection of M documents $D = (\mathbf{d_1}, ..., \mathbf{d_M})$
- **embedding** $X \in \mathbb{R}^{d \times |V|}$, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the embedding of the word at indice i in the vocabulary V and d is the dimension of the embedding

Topic models are probabilistic models used to reveal uncover hidden thematic structures in vast collection of documents. They provide a way to automatically identify topics covered by a text, even abstract ones. One of the most popular topic models nowadays is the Latent Dirichlet Allocation (LDA) model (Blei et al.).

1.2 Latent Dirichlet Allocation

The Latent Dirichlet Allocation model is a 3-layer hierarchical Bayesian model.

The simple idea behind this model is to see documents as distributions over topics and topics as distributions over words. One major assumption before presenting LDA algorithm is that the number of topics in a given corpus is known and fixed. A preprocessing step is conducted to eliminate words that lack semantic significance, often called stopwords in Natural Language Modeing tasks. This involves the removal of frequently occurring words in the documents.

The generative process of LDA is expressed as follow :

Algorithm 1: Generative Process for LDA

Data: Corpus D, Parameters α, β, N
foreach *document* $\mathbf{d}$ *in corpus* D **do**
  Choose $\theta \sim$ Dirichlet(α);
  for w_n *n from 1 to* N **do**
    Choose a topic $z \sim$ Multinomial(θ);
    Choose a word w_n from $p(w_n|z, \beta)$, a multinomial probability conditioned on the topic z;

where the probability density function of the Dirichlet distribution is given by:

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

with $\Gamma(x)$ Gamma function.

The parameters α and β are not known initially. They need to be estimated using variational inference such as the EM algorithm.

1.3 Word Mover Distance

The **Word Mover Distance** (WMD) has been introduced by Kusner et al. overcomes the limitations of traditional distance metrics by explicitly considering the semantic interpretation of words within documents. Unlike LDA, WMD operates at the word level, leveraging pre-trained word embeddings to measure the dissimilarity between documents. It relies on the SkipGram model of Word2Vec Mikolov et al. to embed each words of the document. It results in a matrix $X \in \mathbb{R}^{d \times |V|}$, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the embedding of the word at indice i in the vocabulary V and d is the dimension of the embedding. It is an extension of the common Earth Mover's Metric Rubner et al. where the cost matrix corresponds to the distance between the embeddings of word i and word j, denoted as $c(i,j) = ||\mathbf{x}_i - \mathbf{x}_j||_2$. It introduces document representation as normalized Bag Of Words (nBOW), $\mathbf{d}^i \in \mathbb{R}^|V|$, $i \in \{1, ..., M\}$ where $\mathbf{d}^i_j = \frac{c^i_j}{\sum_{k=1}^n c^i_k}$ and c^i_j is the number of times word j appears in the document $\mathbf{d}^i$.

Then the WMD distance between documents $\mathbf{d}$ and $\mathbf{d}'$ is defined as follow :

$$\min_{\mathbf{T} \geq 0} \sum_{i,j=1}^{|V|} \mathbf{T}_{ij} c(i,j)$$

$$\text{subject to : } \sum_{j=1}^{|V|} \mathbf{T}_{ij} = \mathbf{d}_i, \tag{1}$$

$$\sum_{i=1}^{|V|} \mathbf{T}_{ij} = \mathbf{d'}_j \tag{2}$$

It allows each word i of $\mathbf{d}$ to be transformer, in total or in part, into any word in $\mathbf{d}'$ and the constraints ensure that the totality of each document in transported from one to other.

The complexity associated with the computation of this pairwise distance is $O(|V|log(|V|)$. Thus for corpus with large amount of unique words it may be computationally expensive.

The **Hierarchical Optimal Topic Transport**(HOTT) introduced in this paper is proposed as an alternative to WMD. It leverages the hierarchical structure of LDA while addressing the interpretability and semantic comprehension aspects inherent in word embeddings.

1.4 Related work

Initially, the two most common ways of representing documents were Bag of Word (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF).

However, the features obtained are often not appropriate for computing distances due to their near-orthogonality. Indeed, they reveals little correlations between features and therefore fails to identify synonyms or capture polysemy. In order to prevent this synonymy problem, Latent Semantic Index (LSI)Deerwester et al. has been introduced, applying Singular Value Decomposition to the TF-IDF features to find a latent space that captures most of the variance.

As mentioned previously, topic modeling provides a way to discover the hidden topics of a corpus in a generative way. An alternative to LSI, named probabilistic LSI has been proposed by Hofmann models each document as a mixture of a fixed set of topics, and each word as a sample generated from a single topic. Beyond this, LDA provides a way to initially assign probabilities to unseen documents by introducing a Dirichlet prior.

1.5 Contributions

- **Performances** : HOTT has comparable performances or outperforms WMD and other baselines. It is shown that HOTT has remarkable results on tasks such as KNN-classification and link prediction. While there is no unique best method, HOTT is proven to performs best on average on the KNN-classification task tested over 8 different datasets.

- **Computation cost** : In comparison to WMD that has complexity $O(M^2|V|log(|V|))$ to compute the pairwise distances among all documents in the corpus, HOTT is proved to be significantly efficient.This efficiency stems from its reliance on a hierarchical structure. The initial-level distance, namely the WMD distance between topics, can be precomputed and stored, eliminating the need for individual computations for each document. Moreover, as the second part of the hierarchical structure depends on $|T|$ where $|T| << |V|$, the resulting method is evidently more efficient.

- **Interpretability** : Leveraging topics specific to the corpus, where documents are expressed as distributions over these topics, facilitates a straightforward interpretation of the similarities between documents. This approach provides clarity in understanding the relationships and commonalities among different textual elements in the corpus.

2 Hierarchical Optimal Transport

Hierarchical Optimal Transport Schmitzer and Schnörr (HOT) solves a common problem occuring in dense optimal transport problem. When the number of possible alignments is very large, traditional ways of solving Optimal Transport (OT) problems may become computationally expensive. Let us

consider A and B as the two point clouds that we want to match optimally. The hierarchical approach to OT aims to exploit the structural information of the sets A and B to transform this sets into a collection of clusters or classes.

2.1 General framework

Let us consider a discrete metric space $(\mathcal{X}, d)$, and $\mathcal{P}(\mathcal{X})$ be the space of Borel probability measures on $\mathcal{X}$. We also consider two point clouds X and Y with respectively h and l sub-structures. We define $p = \{p_1, ..., p_h\} \subset \mathcal{P}(\mathcal{X})$ where $p_i \in \Delta^{|X|}$, the probability simplex on $|X|$ elements. We also introduce $q = \{q_1, ..., q_l\} \subset \mathcal{P}(\mathcal{X})$, where $p_i \in \Delta^{|Y|}$, be two sets of probability measures on $\mathcal{P}(\mathcal{X})$. Each probability measures represent structures as distributions over data points. Then, the two collections of structures are defined as $\phi^1 = \sum_{i=1}^{h} \alpha_i \delta_{p_i}$ and $\phi^2 = \sum_{j=1}^{l} \beta_j \delta_{q_j}$ where $\alpha = (\alpha_1, ..., \alpha_h)$ represents the distribution of the set X over the sub-structures and $\beta = (\beta_1, ..., \beta_h)$ represents the distribution of the set Y over the sub-structures. Then the **Hierarchical Optimal Transport problem** reads :

$$HOT(\phi^1, \phi^2) = \min_{\Gamma \in \mathcal{M}_{h \times l}(\mathbb{R}^+)} \sum_{i,j} C_{i,j} \Gamma_{i,j}$$

$$\text{subject to} \sum_{j} \Gamma_{i,j} = \alpha \tag{3}$$

$$\sum_{i} \Gamma_{i,j} = \beta \tag{4}$$

where

$$C_{i,j} = \min_{\mathbf{T} \geq 0} \sum_{k,v} \mathbf{T}_{kv} ||x_k - y_v||_2$$

$$\text{subject to} \sum_{v} \mathbf{T}_{k,v} = p_k \tag{5}$$

$$\sum_{k} \mathbf{T}_{k,v} = p_v \tag{6}$$

Recalling the 1-Wasserstein distance between p and q reads :

$$\mathcal{W}_1(p, q) = \min_{\Gamma \in \mathcal{M}_{h \times l}(\mathbb{R}^+)} \sum_{i,j} C_{i,j} \Gamma_{i,j}$$

$$\text{subject to } \sum_{j} \Gamma_{i,j} = p_i \text{ and } \sum_{i} \Gamma_{i,j} = q_j$$

We can thus rewrite the HOT problem as :

$$HOT(\phi^1, \phi^2) = \mathcal{W}_1 \left(\sum_{i=1}^{h} \alpha_i \delta_{p_i}, \sum_{j=1}^{l} \beta_j \delta_{q_j} \right)$$

where the cost metric is defined as $(\mathcal{W}_1(p_i, q_j))_{i \in \{1,..,h\} \text{ and } j \in \{1,...,l\}}$

2.2 Hierarchical Optimal Topic Transport

We aim to use the advantages of Hierarchical Optimal Transport (HOT) to compute distances between documents in a corpus D consisting of M documents. The interpretability of the HOTT model is achieved by incorporating word embeddings from the Word2Vec model, represented by the matrix $X \in \mathbb{R}^{d \times |V|}$ where d and $|V|$ are respectively the dimensions of the embedding and of the vocabulary. Then, using the LDA model, we are able to produce a set of corpus-specific topics distribution over word embeddings, denoted as $T = \{t_1, t_2, ..., t_{|T|}\} \in \Delta^{|V|}$. It also allows to compute document distributions over topics $\bar{d^i} \in \Delta^{|T|}$

Then, the Hierarchical Optimal Topic Transport problem between documents d_1 and d_2 reads :

$$HOTT(d^1, d^2) = \mathcal{W}_1 \left(\sum_{k=1}^{|T|} \bar{d_k^1} \delta_{t_k}, \sum_{k=1}^{|T|} \bar{d_k^2} \delta_{t_k} \right)$$

where δ_{t_k} corresponds to the Dirac on the corresponding topic t_k and the ground metric is defined as $(WMD(t_i, t_j))_{i,j \in \{1,...,|T|\}}$. One can notice that the WMD distance in this case is the 1-Wasserstein distance with cost matrix having entries $C_{ij} = ||x_i - x_j||_2$.

The hierarchical structure arises from the computation of two distinct optimal transport problems. Initially, we calculate the 1-Wasserstein distance between topic distributions, treating it as the first optimal transport problem. Subsequently, we compute the 1-Wasserstein distance between document distributions over topics, utilizing the first computed matrix as the cost matrix for the second transport problem.

2.3 Theoretical guarantees

Proof of HOTT being a metric. The metric HOTT relies on the distinctiveness of document distributions across various topics. The fundamental premises are that every document can be expressed as an unique linear combination of topics, denoted as $d^i = \sum k = 1^{|T|} \bar{d_k^i} t_k$ and that each topic is inherently distinct. This assumptions ensure that $HOTT(d^i, d^j) = 0 \iff d^i = d^j$. Then the positivity, symmetry and the triangle inequality follow from the properties of $\mathcal{W}_1$.

Truncation. If each word corresponds to a single topic, ie. $|T| = |V|$, the HOTT metric becomes the WMD. Allowing for variability in $|T|$ introduces different levels of granularity in defining topics. This flexibility presents a trade-off between computational speed and the specificity of the identified topics. In practice, the distribution of topic proportions within documents tends to be dense, with little mass outside of the dominant topics. However, in topic models, one assumption is that documents are assumed to be represented by a small number of topics. Then to compute HOTT, we truncate LDA topic proportions at $1/|T| + 1$ to force sparsity of document's distribution over topics. This allows to greatly improve computational efficiency.

Complexity The 1-Wasserstein distance can be solved with complexity $O(l^3 log(l))$ using the Hungarian algorithm by Kuhn, with $l = max(n, m)$. One common way to improve computational performance, reducing to $O(l^2/\epsilon^2)$ is using the *Entropic Regularization* for discreate measures. It employs the Kullback-Leibler divergence as regularizing function to obtain approximate solution to the original transport problem. The Kullback-Leibler divergence is defined as

$$KL(P||Q) = \sum_{i,j} P_{i,j} log\left(\frac{P_{i,j}}{Q_{i,j}}\right) - P_{i,j} + Q_{i,j}$$

Then, the Entropic Regularization of the OT problem reads :

$$L_C^{\epsilon}(a, b) = \min_{P \in U(a,b)} \langle P, C \rangle + \epsilon KL(P||a \otimes b)$$

However, as the number of topics is deliberately small, the support of our distribution is not sufficiently large to allow the Enropic Regularization to perform well. Therefore, linear programming would be preferred and faster if we need an accurate solution.

The efficiency of the HOTT metric relies on the hierarchical structure. The first-level distance namely the distance between topics can be precomputed and stored. On the contrary, computing pairwise metric, like WMD, between two documents implies to iterate over all the vocabulary, as documents as considered as normalized BOW. Thus the complexity for computing pairwise distances of an entires corpus of M documents amounts to $O(M^2|V|log(|V|)$. There is clearly a trade-off between the precision of the metric which implies large amount of unique words and the computational cost. Assuming the distance between topics is already precomputed, the complexity of HOTT relies on $O(|T|^2 lof(|T|)$. As $|T| << |V|$, the HOTT wil prove to be signicantly efficient.

2.4 Numeric comparisons : Exploring differences between Hierarchical Optimal Transport distance and other Hierarchical distances

We aim to investigate alternative hierarchical distances and assess their performance in comparison to the HOT distance. To achieve this, we conduct a comparative analysis using a straightforward classification task centered on sports articles, as outlined in Yurochkin et al..

Note. The primary focus here lies in comparing performance metrics rather than delving extensively into the interpretability of the NLP classification task itself.

Dataset. Our dataset consists in 737 sports news articles from the *BBC Sport* in five topical areas from 2004-2005. A number from 0 to 4 is attributed to each document aligning with its topic label including *athletics, cricket, football, rugby tennis*. We perform word embeddings using Global Vector representation (GloVe) Pennington et al. with 6 Billion tokens and 300 features. We then reduce the size of the vocabulary by stemming and removing stopwords.

We will consider two ways of computing the topics and documents distributions.

- For the first one, we leverage the LDA model to simulate 30 topics as distributions over words and documents as distributions over the induced topics. Upon obtaining topics histograms over words, $t_k \in \mathbb{R}^v \forall k = 1, ...30$ where v is the length of the vocabulary after reduction and histograms characterizing topics as distributions for documents $d_k \in \mathbb{R}^{30}$, we can calculate hierarchical distances between these structured distributions.

- The second way will use sparse vectors. Indeed, when dealing with distances between histograms, we can often face sparsity. A common example will be with temperature histograms over months. Temperatures reach in the summer would not be achieved in the winter, resulting in 0 probability of having extremely hot values in winted and vice-versa. For our case, we are truncating each topic probability vector to their top-20 words, assigning 0 values to the others.

Classification task. We evaluate our distance on k-Nearest Neighbors classification, with $k = 3$.

Additionally to HOTT metric, we are evaluating and comparing several other hierarchical metrics described below :

L1-OT distance. We calculate the L1-distance between topics histograms and subsequently then the 1-Wasserstein distance between documents. The cost matrix for the 1-Wasserstein distance is determined by the previously computed L1-distances between topic histograms.

KL-OT distance The Kullback-Leibler divergence or relative entropy, denoted $KL(P||Q)$ is a statistical measure of similarity between two measures that we have introduced before. We calculate the KL-distance between topics histograms and subsquently then the 1-Wasserstein distance between documents. The cost matrix for the 1-Wasserstein distance is determined by the previously computed L1-distances between topic histograms.

2.4.1 Results for dense vectors

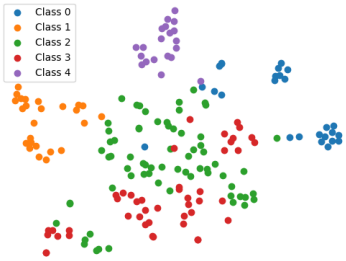

Figure 1: True labels repartition

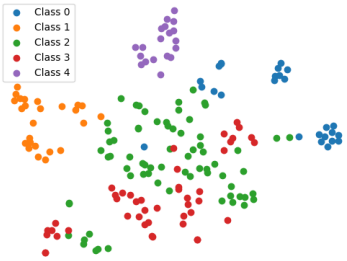

Figure 2: Predicted labels induced by HOTT metric

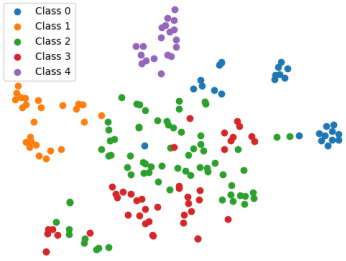

Figure 3: Predicted labels induced by the L1-OT metric

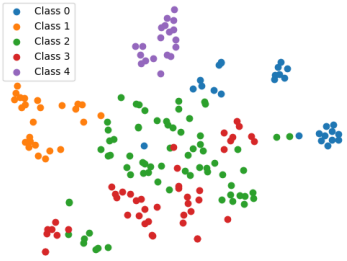

Figure 4: Predicted labels induced by the KL-OT metric

We observe that the choice of the ground metric to compute $\mathcal{W}_1$ distance between documents does not influence the overall performance of metric. The accuracy stays very high while using KL-OT distance and L1-OT distance. Those observations are rather surprising since we would have expect the choice of the cost matrix to have a significant influence of the overall performances.

Table 1: Comparison of the KNN classification for induced by different metrics

Metric	Accuracy (%)	Test error
HOTT	95.14	0.048649
L1-OT	95.03	0.048621
KL-OT	94.59	0.054054

2.4.2 Results for sparse vectors

We perform the exact same experiment takings the sparse topic vectors.

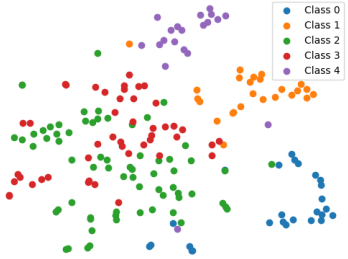

Figure 5: True labels repartition

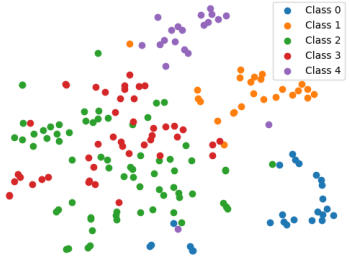

Figure 6: Predicted labels induced by HOTT metric

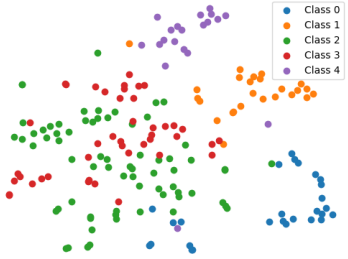

Figure 7: Predicted labels induced by the L1-OT metric

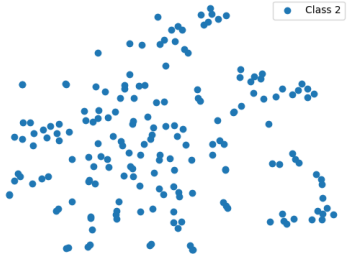

Figure 8: Predicted labels induced by the KL-OT metric

As we would have expected, the Kullback-Leibler divergence fails to capture similarities between distributions when dealing with sparse entries, due to the logarithm. However the L1-OT metric still have competitive result on our task.

Another approach is to use the Maximum Mean Discrepancies metric (MMD). MMD is a kernel statistical method used to determine the similarity

Table 2: Comparison of the KNN classification for induced by different metrics for sparse topics vectors

Metric	Accuracy (%)	Test error
HOTT	97.84	0.021622
L1-OT	96.76	0.064865
KL-OT	0.36756	0.054054

between two distributions.

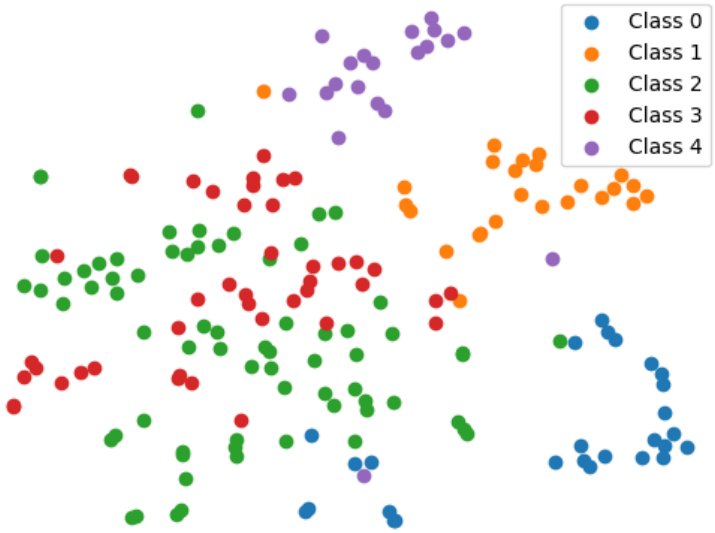

Figure 9: Predicted labels with the MMD-OT metric

3 Conclusion

We have performed a k-NN classification using several hierarchical metrics, HOTT, L1-OT, KL-OT and MMD-OT. The results showed that the the ground metric used to performed the HOT metric does not really influence the result of the classification. However, when dealing with sparse histograms, some metrics as MMD and KL are totally inappropriate. One key element in the choice of a good hierarchical metric is that the HOT metric takes into account the distance between the initial input data.In our usecase, the HOTT metric is first calculating the 1-Wasserstein distance be-

tween topics taking the L1-distance between embeddings as cost matrix. On the other hand, L1-OT metric or KL-OT metric are only considering distance between the topics histograms. When dealing with NLP tasks such as documents classification or summarization, considering words embeddings will beneficial to ensure interpretability and preserve semantic. Going further, it would be interesting to look at the resistance to noise of all those hierarchical metrics.

3.1 Perspectives

The HOTT has proven to perform significantly well on classification task, providing competitive computation cost as well as interpretability. An interesting perspective to go further would be to include document structure to the model. Certain type of documents as research papers are organized in sections. Some words occurring in the *Related Work* section and words appearing the *Method* section could be likely to have close semantic, but need to be distinguished in the research paper. Thus adding a penalty term acting as a regularization for words occurring in different pats of the document would be interesting to try.

4 Link to the Computational Optimal Transport course

The central part of the paper Yurochkin et al. is the discrete-to-discrete optimal transport problem, or Kantorovitch problem :

$$L_c(a,b) = \min_{P \in U(a,b)} \langle P, C \rangle$$

where

$$U(a,b) = \{P \in \mathbb{R}_+^{n \times m} : P\mathbb{1}_m = a \text{ and } P^t\mathbb{1}_n = b\}$$

The hierarchical distance presented allows to compute more efficient the Kantorovitch problem assuming that we can admit clusters/structures in the input data and represent input data as histograms over the clusters.

The proof of $\mathcal{W}_1$ being a metric which is crucial for the introduction of the HOTT metric is part of the course.

Proof that $W_p(a,b) = L_{D^p}(a,b)^{\frac{1}{p}}$ defines a distance :

$C = D^p$ has a null diagonal, $W_p(a,a) = 0$, with corresponding optimal transport matrix $P^\star = \text{diag}(a)$; by the positivity of all off-diagonal elements of $D^p, W_p(a,b) > 0$whenever $a \neq b$ (because in this case, an admissible coupling necessarily has a non-zero element outside the diagonal); by symmetry of $D^p, W_p(a,b) = 0$is itself a symmetric function.

To prove the triangle inequality of Wasserstein distances for arbitrary measures,uses the gluing lemma, which stresses the existence of couplings with a prescribed structure. In the discrete setting, the explicit construction of this glued coupling is simple. Let $a, b, c \in \Sigma^n$. Let $P$and Q be two optimal solutions of the transport problems between a and b, and b and c respectively. We define $b_j$if $b_j > 0$ and set otherwise $\overline{b}_j = 1$(or actually any other value). We then define

$$S = P\text{diag}(1/b)Q$$

We remark that $S \in U(a, c)$ because

$$S1_n = P\text{diag}\begin{pmatrix}\frac{1}{\overline{b}}\end{pmatrix} Q1_n = P\begin{pmatrix}\frac{b}{\overline{b}}\end{pmatrix} = P1_{\text{Supp}(b)} = a$$

References

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. 3(3): 993–1022.

Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. 41(6):391–407.

T. Hofmann. Probabilistic latent semantic indexing. page 350–57.

H. W. Kuhn. The hungarian method for the assignment problem. 2:83–97.

M. J. Kusner, N. I. K. Yu Sun, and K. Q. Weinberger. From word embeddings to document distances. 37:957–966.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. page 1532–1543.

Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases.

B. Schmitzer and C. Schnörr. A hierarchical approach to optimal transport.

M. Yurochkin, S. Claici, E.Chien, F. Mirzazadeh, and J. Solomon. Hierarchical optimal transport for document representation.