Optimal Transport Geometry for Sentiment Analysis

Xuedong Shang[1]
Supervised by: **Antoine Rolet**[2] and **Marco Cuturi**[3]
From 23 May to 29 July

[1] Ecole normale supérieure de Rennes, Bruz, France
xuedong.shang@ens-rennes.fr
[2] Graduate School of Informatics, Kyoto University, Kyoto, Japan
antoine.rolet@iip.ist.i.kyoto-u.ac.jp
[3] Yamamoto-Cuturi Lab., Kyoto University, Kyoto, Japan
mcuturi@i.kyoto-u.ac.jp

Abstract. This report introduces a regression model for modeling the relationship between histogram-valued data and a real-valued parameter. It uses the theory of optimal transport, and especially the notion of Wasserstein barycenter based on Wasserstein distance. We try to apply this technique to a real-life problem which is the sentiment analysis problem.

Keywords: optimal transport, earth mover's distance, Sinkhorn distance, bag-of-words model, Wasserstein barycenter, sentiment analysis, geodesic regression

1 Introduction

Several research opportunities exist at the interaction of machine learning and optimal transport (OT), as demonstrated by recent works carried out in Yamamoto-Cuturi Laboratory. The goal of this nascent field is to use the optimal transport geometry on probability distributions in a machine learning context, where probability distributions appear under the form of histograms of features and statistical models.

Probability histograms play an crucial role in many different fields of computer science, such as graphics, natural language processing, etc. The feature space on which these probability distributions are supported can often be endowed with a distance.

Optimal transport theory [14] proposes a natural way to define a distance between probability histograms on features using a distance between features. Intuitively, it sees probability histograms as different ways of piling up a certain amount of dirt and quantifies the distance between two of them as the minimum cost of turning one pile into the other. Thus, two histograms having only a small difference in the optimal transport sense could be very dissimilar under usual metrics such as $\mathbb{L}_1$, $\mathbb{L}_2$, etc.

Since the physical interpretation of optimal transport distance, a.k.a.Wasserstein distance, is totally different from usual metric, it also defines extremely different ways to interpolate between histograms. These optimal transport interpolations, a.k.a.Wasserstein barycenters, maybe particularly useful in histogram-valued regression problem.

In a practical perspective, these interpolations can only find applications if they are computationally tolerant. While direct approaches to carry out these interpolations have been proven to be intractable by Agueh and Carlier [9], some regularized versions have been proposed [2][10] to provide cheap computational cost which makes them useful in real-life problems.

In this paper, we propose a Wasserstein distance-based regression model, a.k.a.Wasserstein geodesic regression, on histogram-valued inputs, and try to apply it on the sentiment analysis problem. That is to say, given a review text on some kind of product, we try to predict the attitude of the writer towards this product.

We start this report with some reminders on transportation theory and Wasserstein distance in section 2. Then we propose a Wasserstein distance-based regression method on histogram-valued data in section 3. Finally we try to deal with the sentiment analysis problem using this regression model before concluding.

2 Background

2.1 Optimal Transport

A *transportation problem* [1] is the study of optimal transportation and allocation of resources. For example, suppose that we have several suppliers, each with a given amount of goods. They are required to supply several consumers, each with a given limited capacity. For each supplier-consumer pair, the cost of shipping a single unit of supplies is known. The problem is then to determine a flow to ship the supplies in order to minimize the total cost of transportation.

This is indeed a bipartite network flow problem which can be formalized as a linear programming problem. Let I be the set of suppliers, J be the set of consumers and $M = [m_{ij}]$ be the matrix of cost of shipment. Let x_i be the total supply of the supplier i and y_j be the total capacity of consumer j. We want to find a flow $F = [f_{ij}]$ to minimize the overall cost, and the linear programming problem associated can be thus formulated as,

$$\begin{array}{lll} \textbf{min.} & \sum_{i\in I}\sum_{j\in J} m_{ij}f_{ij} & \\ \textbf{s.t.} & f_{ij} \geq 0 & i \in I, j \in J \\ & \sum_{i\in I} f_{ij} = y_j & j \in J \\ & \sum_{j\in J} f_{ij} \leq x_i & i \in I. \end{array} \tag{1}$$

The first constraint here allows only shipment from suppliers to consumer, the second one forces the consumers to fill up all of their capacities and the last one limits the supply sent by each supplier to its total amount.

Remark 1. A reasonable condition is that we can always assume that $\sum_{j\in J} y_j \leq \sum_{i\in I} x_i$, since we can always switch the role of what we call suppliers and what we call consumers.

2.2 Earth Mover's Distance

Feature distributions are widely used in many computer science domains. For example, in image retrieval, we often have to deal with the one-dimensional distribution of image intensities and three-dimensional distribution of image colors. It seems very important to define a distance between distributions so as to cope with such problems . The EMD [4] is a measure to evaluate dissimilarity between two multi-dimensional distributions in some feature space. A distance measure between single features, called *ground distance*, is required and has to be chosen by hand for each specific problem (it can be any distance in general).

In fact, a distribution can be considered as a set of clusters where each cluster is represented by its mean or mode and its weight (the fraction of the distribution that belongs to that cluster). Such a representation is called the *signature* of the distribution. The transportation problem, which is presented previously, can be naturally used for signature matching by defining one signature as suppliers and other signature as consumers. Thus, the computation of the EMD can be based on the solution to the transportation problem.

For instance, consider two signatures $\mathbf{p} = \{(p_1, w_{p_1}), \cdots, (p_m, w_{p_m})\}$, where p_i is the cluster representative and w_{p_i} is the weight of the cluster, and $\mathbf{q} = \{(q_1, w_{q_1}), \cdots, (q_n, w_{q_n})\}$. Suppose that the ground distance matrix/cost matrix M is also given. We want to find a flow F that minimize the overall cost. And we can formulate this linear programming problem just like the transportation problem where w_{p_i} corresponds to the supply x_i and w_{q_j} corresponds to the capacity y_j in the previous subsection. We can then define the EMD.

Definition 1 (Earth Mover's Distance).

$$EMD(\boldsymbol{p}, \boldsymbol{q}) \coloneqq \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{j=1}^{n} y_j} \tag{2}$$

where the denominator is a normalization factor to avoid favoring signatures with smaller total weights.

Remark 2. The EMD is a true distance when the compared two distributions have the same overall mass (integral). In that case, it is equivalent to the 1st *Wasserstein distance* between two distributions.

The EMD extends naturally the notion of distance between single elements to distance between distributions. It can be applied to more general *variable-size* signatures, thus to histograms of features, contrary to other distances like $\mathbb{L}_1$ proposed in [5], $\mathbb{L}_2$ or $\mathbb{L}_\infty$ proposed in [6] and a weighted version of $\mathbb{L}_2$ proposed in [7].

2.3 Sinkhorn Distance

The EMD defines a more powerful geometry to compare probabilities at the cost of a heavy computational effort (at least in $O(d^3 \log(d))$ where d is the dimension of the histograms/signatures). The Sinkhorn distance is proposed by [2] to overcome this problem.

Indeed, we can look at the optimal transport problem in an entropic perspective. Consider two probability vectors $\mathbf{p}$ and $\mathbf{q}$ in the simplex $\Sigma_d \coloneqq \{\mathbf{x} \in \mathbb{R}_+^d \mid \mathbf{x}^T \mathbb{1}_d = 1\}$ where $\mathbb{1}_d$ is the d-dimensional vector of ones. We can define the so-called *transport polytope* of $\mathbf{p}$ and $\mathbf{q}$,

$$U(\mathbf{p}, \mathbf{q}) \coloneqq \{T \in \mathbb{R}_+^{d \times d} \mid T\mathbb{1}_d = \mathbf{p}, T^T \mathbb{1}_d = \mathbf{q}\}. \tag{3}$$

Remark 3. $U(\mathbf{p}, \mathbf{q})$ contains indeed all possible joint probabilities of (X, Y) that take values in $\{1, \cdots, d\}$.

Given the ground distance matrix M, we can thus reformulate the optimal transport problem as,

$$d_M(\mathbf{p}, \mathbf{q}) \coloneqq \min_{T \in U(\mathbf{p}, \mathbf{q})} \langle T, M \rangle \tag{4}$$

where $\langle \cdot, \cdot \rangle$ is the dot-product.

Definition 2 (Sinkhorn Distance).

$$d_{M,\alpha}(\boldsymbol{p}, \boldsymbol{q}) \coloneqq \min_{T \in U_\alpha(\boldsymbol{p}, \boldsymbol{q})} \langle T, M \rangle \tag{5}$$

We compute in practice a regularized term called *dual-Sinkhorn divergence* which is defined for all $\lambda > 0$,

$$d_M^\lambda(\mathbf{p}, \mathbf{q}) \coloneqq \left\langle T^\lambda, M \right\rangle \tag{6}$$

where $T^\lambda = \arg\min_{T \in U(\mathbf{p}, \mathbf{q})} \langle T, M \rangle - \frac{1}{\lambda} h(T)$.

2.4 Regularized Wasserstein Barycenters

An important notion to be determined in problems like histogram or probability distribution regression is how to calculate the histograms barycenter.

With the previous definition, we can define the so called *entropy regularized OT distance* or *regularized Wasserstein distance* between two histograms $\mathbf{p}$ and $\mathbf{q}$ as,

$$W(\mathbf{p}, \mathbf{q}) \coloneqq \min_{T \in U(\mathbf{p}, \mathbf{q})} \{\left\langle T^\lambda, M \right\rangle + \gamma H(T)\} \tag{7}$$

where $H(T)$ is the negative entropy of T (contrary to $h(T)$ in the previous subsection which is the entropy of T), and thus γ is a positive regularization parameter.

Following Agueh and Carlier's definition of *Wasserstein Barycenter* [9] as Fréchet means in a Wasserstein space, Cuturi and Doucet proposed in [8] a way of computing the barycenter of s histograms $(\mathbf{p}_s)_s$ using the regularized Wasserstein distance.

Definition 3 (Regularized Wasserstein Barycenter). *The barycentric map $P : \Sigma_s \to \Sigma_d$ that associates to a vector $\lambda \in \Sigma_s$ the barycenter of $(\boldsymbol{p}_s)_s$ with weights λ is uniquely defined as,*

$$P(\lambda) \coloneqq \underset{\boldsymbol{p} \in \Sigma_d}{\arg\min} \sum_s \lambda_s W(\boldsymbol{p}, \boldsymbol{p}_s). \tag{8}$$

Remark 4. The uniqueness mentioned in the definition just above comes from the strong convexity of the right-hand side of the equation.

3 Wasserstein Geodesic Regression

Now we come to our contributions in this section. Assume that we are given a dataset $((t_1, \mu_1), \cdots, (t_n, \mu_n))$ where t_i are real values and μ_i are probability distributions or histograms. Our objective is to find a geodesic g_t which best describes this dataset, i.e., intuitively, a geodesic such that each g_{t_i} is close to μ_i. Here we will present two possible parametrizations of g_t.

3.1 Lagrangian Approach

In this first approach, we suppose that μ_i are probability distributions on $\mathbb{R}^d$ and we use the squared Euclidean distance as ground metric. Following Seguy and Cuturi's approach [11], we consider generalized geodesics parametrized through a basis probability measure σ on $\mathbb{R}^d$ and two velocity fields V_1 and V_2 in $L^2(\sigma, \mathbb{R}^d)$ defined on the support of σ,

$$g_t^\alpha = (\mathrm{id} - V_1 + t(V_1 + V_2))\#\sigma. \tag{9}$$

The geodesic regression problem consists thus in minimizing,

$$\sum_{i=1}^{n} W^2(\mu_i, g_{t_i}^\alpha) \tag{10}$$

over σ, V_1 and V_2.

Remark 5. We will not use this first approach in our experiments, we will only use the method to be introduced in the next subsection.

3.2 Eulerian Approach

A more general approach that allows us to generalize to more cost functions other than the square Euclidean distance and more general ground spaces other than $\mathbb{R}^d$. However, in this case, we will not have a closed formula for the parametrization as in the previous approach. We define a geodesic between two end measures ν and η as the solutions of the convex problem,

$$g_t(\nu, \eta) = \underset{\mu}{\arg\min}(1-t)W(\mu, \nu) + tW(\mu, \eta) \tag{11}$$

which refers to the Wasserstein barycenter problem [9] of two measures.

The geodesic regression problem consists in minimizing,

$$\sum_{i=1}^{n} W^2(\mu_i, g_{t_i}(\nu, \eta)) \tag{12}$$

over ν and η.

3.3 Automatic Differentiation

In order to minimize the quantity in (12), we need to build a numerical scheme upon gradient descent. Intuitively, we need to find a geodesic that best fits our training set by adjusting the position of the two end measures.

Let's define $\varepsilon(\nu, \eta) \coloneqq W(\mu, g_t(\nu, \eta))$, our goal is to compute the optimal solution to problem,

$$\arg\min_{\nu,\eta} \varepsilon(\nu, \eta). \tag{13}$$

The energy of this problem is generally not convex. We thus have to recover a stationary point of that energy through gradient descent. The gradient of ε with respect to ν and η can be computed using the chain rule:

$$\nabla\varepsilon_\nu(\nu, \eta) = [\partial_\nu g_t(\nu, \eta)]\top\nabla W(\mu, g_t(\nu, \eta)), \tag{14}$$

$$\nabla\varepsilon_\eta(\nu, \eta) = [\partial_\eta g_t(\nu, \eta)]\top\nabla W(\mu, g_t(\nu, \eta)) \tag{15}$$

where $\partial_\nu g_t(\nu, \eta)$ is the Jacobian of $\nu \mapsto g_t(\nu, \eta)$, $\partial_\eta g_t(\nu, \eta)$ is the Jacobian of $\eta \mapsto g_t(\nu, \eta)$ and $\nabla W(p, q)$ is the gradient of the loss $q \mapsto W(p, q)$.

The gradient of $W(p, q)$ can be computed as,

$$\nabla W(p, q) = \gamma \log(a) \tag{16}$$

where $a \in \mathbb{R}^d$ is the left scaling produced by Sinkhorn's fixed-point algorithm.

And finally, since the exact computation of the Jacobian above is impratical, we can consider using an automatic differentiation approach.

4 Experiments

Sentiment analysis is a process of computationally identifying opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. In this section, we try to deal with this real-life problem using the Eulerian method we proposed in the previous section.

More precisely, our regression input will be some review texts and we want to predict the score (from 1 to 5) of a given text using our model.

Dataset and Pre-processing The dataset we use for sentiment analysis is the review texts on laptops from *Amazon.com*. One can find these data on `http://times.cs.uiuc.edu/~wang296/Data/`. And the main part of our implementations is written in matlab.

Firstly, we need to parse these data into a suitable format including only review texts and scores. We made two *mat* files, one with review texts and anther one with scores in columns. We also preserved their order with respect to each other.

Next, we need to convert these texts into bag-of-words. In this part, we need to maintain the count for every word since later we need to normalize these bag-of-words into probability histograms.

Remark 6. The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity.

Word2vec Models Now, a *Word2vec* model is required to produce word embeddings for our data. Word2vec takes as its input a large corpus of text and produces a high-dimensional space, with each unique word in the corpus being assigned a corresponding vector in the space so that we can compute ground distance between different review texts.

What we use in our experiments is a pre-trained Word2vec model which can be found on `https://code.google.com/archive/p/word2vec/`. This model is trained on part of Google News dataset (about 100 billion words). The model contains 300-dimensional vectors for 3 million words and phrases. The phrases were obtained using a simple data-driven approach described in [3].

With this pre-trained Word2vec model at hand, we can now create our own dictionary. To do so, we need to create the set of words in the entire data. This time, we do not need to keep the count for every single word. However, we can remove all the punctuations and stop words as they are not meaningful in the sentiment analysis sense. This step can also reduce significantly the size of our training and testing data.

We can pass this dictionary to the pre-trained Word2vec model to obtain real training and testing data. This part of implementations are encoded in C++. Then we are ready to pre-compute the ground distance matrix. In our implementations, we used two different distances: Euclidean distance and cosine distance. The pre-computation of this distance matrix can largely accelerate our experiments.

First Experiments Now that ground distance matrix has been computed, we are ready to launch our first experiments. We first compute the Wasserstein barycenter of texts with score 1 and texts with score 5. We obtain thus two end measures denoted respectively ν and η. In these first experiments, we compute our geodesic between these two end measure by discretizing it. Indeed, in our

experiments, we choose t from 0 to 1 with a step of 0.025, then we compute for each t, the solution of the convex problem (11). Then we need to project our review texts on this geodesic to compute the overall error (12).

We got very poor results from these first experiments, almost 96% of review texts are projected on the two end measures as we can see from Fig. 1 to Fig. 5. Indeed, Fig. 1 represents distances between 20 randomly selected review texts with score 1 and every discretized points on the geodesic, so as for Fig. 2 to Fig. 5. We can see that texts of score 1 are mostly closer to the end measure of score 1 and texts of score 5 are mostly closer to the end measure of score 5. However, the evolution curve of texts of score 2, 3 and 4 are somehow too flat.

Next Step Since we gained very poor results from our first experiments, we can now carry out the gradient descent scheme on the end measures as we mentioned in the previous section. To do so, we need to implement an automatic differentiation algorithm. This can be done using an existing library. Unfortunately, my code is not totally compatible with these libraries, and I did not succeed in rewriting my code finally.

5 Conclusion

In this report, we proposed a Wassertein geodesic regression model so as to predict the writers' attitude based on their review texts.

Performance The first results which only used a discretized geodesic between two end measures seem to be very disappointing. A reasonable explanation of the poor performance maybe the fact that we do not have a correct ground metric. As we can see in Fig. 6, the cosine distance is not really meaningful for our review texts.

Then I did not manage to implement a correct version of automatic differentiation of the Jacobian. As a result, I was not able to carry out the numerical scheme to adjust the geodesic which builds upon gradient descent, and thus required the computation of the high-dimensional Jacobian of the barycenter operators $\nu \mapsto g_t(\nu, \eta)$ and $\eta \mapsto g_t(\nu, \eta)$.

Future Work As we mentioned above, using simply an usual ground metric such as cosine distance or Euclidean distance may not be appropriate for our problem. Thus a metric learning process on the ground metric can be conceivable.

Secondly, a proper implementation of algorithmic differentiation of the Jacobian is required. With this tool at hand, we can then run real experiments using gradient descent.

More tests on different kind of datasets may also be a good idea to see the performance of our approach afterwards.

We talked a little about another approach which is the Lagrangian approach. This can also be a possible way of solving our problem.

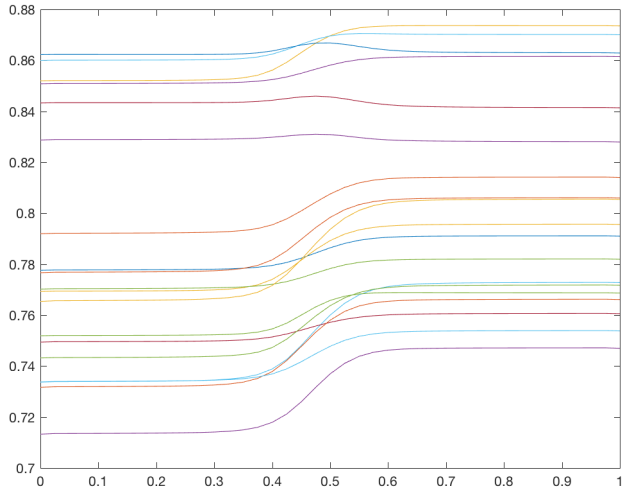

Fig. 1. Distance between random selected review texts with score 1 and the geodesic

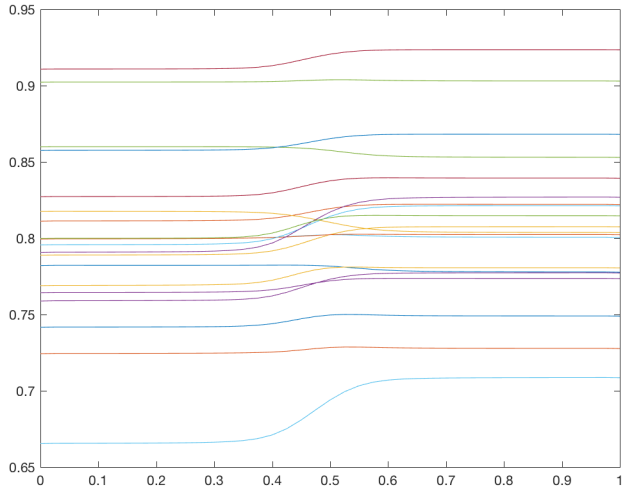

Fig. 2. Distance between random selected review texts with score 2 and the geodesic

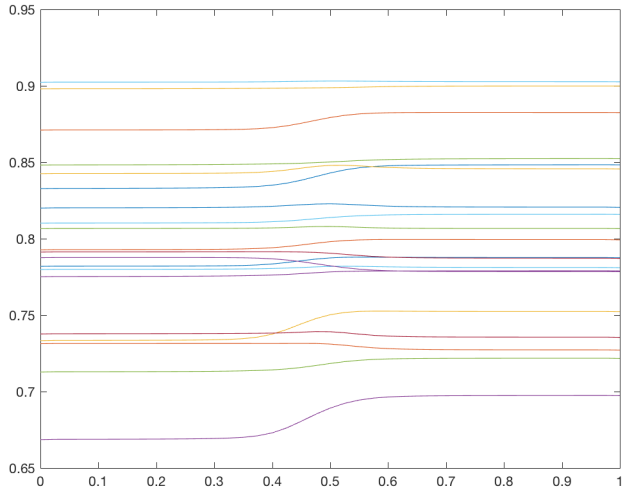

Fig. 3. Distance between random selected review texts with score 3 and the geodesic

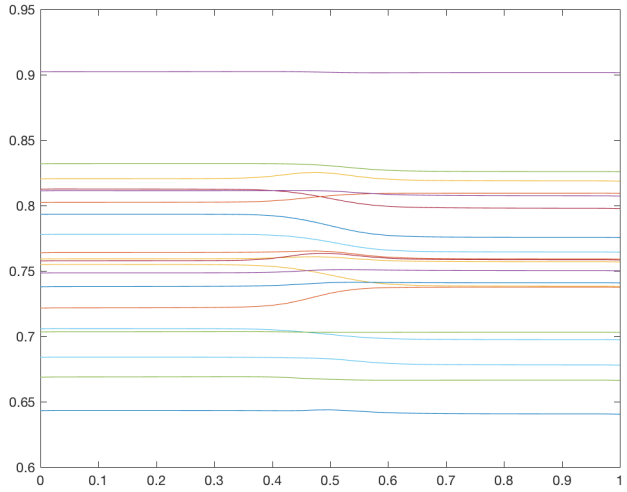

Fig. 4. Distance between random selected review texts with score 4 and the geodesic

Acknowledgements. The internship opportunity I had at Kyoto University was a great chance for learning and professional development. Therefore, I consider myself as a very lucky individual as I was provided with an opportunity to be a part of it. I am also grateful for having a chance to meet so many wonderful people and professionals who led me though this internship period.

I express my deepest thanks to my supervisor Marco Cuturi for taking part in useful decision and giving necessary advices and guidance and arranged all facilities to make life easier. I choose this moment to acknowledge his contribution gratefully.

I would also like to express my best regards, deepest sense of gratitude to Antoine Rolet, Vivien Seguy for their careful and precious guidance which were extremely valuable for my study both theoretically and practically.

References

1. Hitchcock, Frank L.. The distribution of a product from several sources to numerous localities. Journal of mathematics and physics 20.1, 224-230 (1941)
2. Cuturi, Marco. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in Neural Information Processing Systems (2013)
3. Mikolov, T., and J. Dean. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems (2013)
4. Rubner, Yossi, Carlo Tomasi, and Leonidas J. Guibas. A metric for distributions with applications to image databases. Computer Vision, 1998. Sixth International Conference on IEEE (1998)
5. Swain, Michael J., and Dana H. Ballard. Color indexing. International Journal of Computer Vision 7.1, 11-32 (1991)
6. Stricker, Markus A., and Markus Orengo. Similarity of color images. IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology. International Society for Optics and Photonics (1995)
7. Niblack, Carlton W., et al.. QBIC project: querying images by content, using color, texture, and shape. IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology. International Society for Optics and Photonics (1993)
8. Cuturi, Marco, and Arnaud Doucet. Fast computation of Wasserstein barycenters. arXiv Preprint arXiv:1310.4375 (2013)
9. Agueh, Martial, and Guillaume Carlier. Barycenters in the Wasserstein space. SIAM Journal on Mathematical Analysis 43.2: 904-924 (2011)
10. Benamou, Jean-David, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. SIAM Journal on Scientific Computing 37.2: A1111-A1138 (2015)
11. Seguy, Vivien, and Marco Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. Advances in Neural Information Processing Systems (2015)
12. Bonneel Nicolas, Gabriel Peyré, and Marco Cuturi. Wasserstein barycentric coordinates: histogram regression using optimal transport. ACM Transactions on Graphics 35.4 (2016)
13. Hunter, David R., and Kenneth Lange. Quantile regression via an mm algorithm. Journal of Computational and Graphical Statistics 9.1: 60-77 (2000)
14. Villani, Cédric. Optimal transport: old and new. Vol. 338. Springer Science & Business Media (2008)

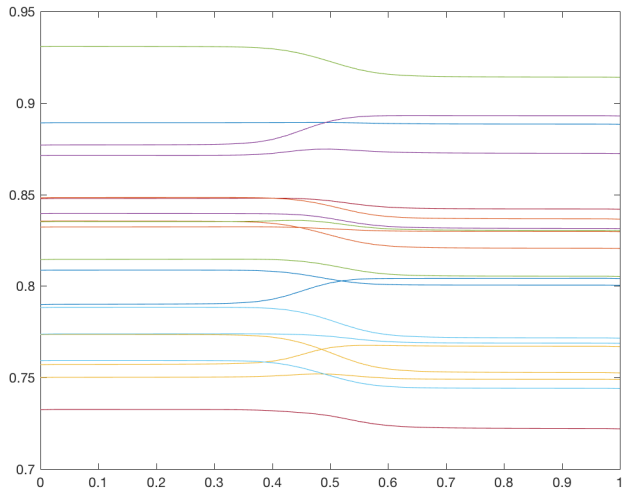

Fig. 5. Distance between random selected review texts with score 5 and the geodesic

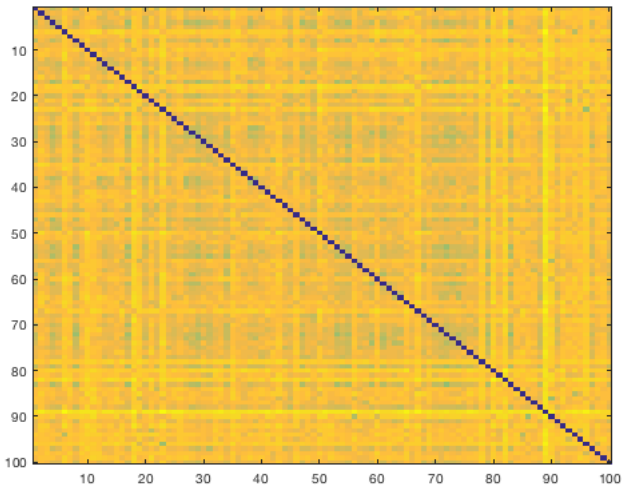

Fig. 6. Ground distances between randomly selected review texts, the first 20 texts are texts of score 1, the next 20 texts are texts of score 2, etc.