Proceedings of Machine Learning Research 129:721–736, 2020 ACML 2020

Robust Document Distance with Wasserstein-Fisher-Rao Metric

Zihao Wang wzh17@mails.tsinghua.edu.cn
BNRist, Department of Computer Science and Technology, RIIT, Institute of Internet Industry, Tsinghua University

Datong Zhou zdt14@mails.tsinghua.edu.cn
Department of Mathematical Sciences, Tsinghua University

Ming Yang yang-m17@mails.tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University

Yong Zhang zhangyong05@tsinghua.edu.cn
BNRist, Department of Computer Science and Technology, RIIT, Institute of Internet Industry, Tsinghua University

Chenglong Bao clbao@mail.tsinghua.edu.cn
Yau Mathematical Sciences Center, Tsinghua University

Hao Wu hwu@tsinghua.edu.cn
Department of Mathematical Sciences, Tsinghua University

Editors: Sinno Jialin Pan and Masashi Sugiyama

Abstract

Computing the distance among linguistic objects is an essential problem in natural language processing. The word mover's distance (WMD) has been successfully applied to measure the document distance by synthesizing the low-level word similarity with the framework of optimal transport (OT). However, due to the global transportation nature of OT, the WMD may overestimate the semantic dissimilarity when documents contain unequal semantic details. In this paper, we propose to address this overestimation issue with a novel Wasserstein-Fisher-Rao (WFR) document distance grounded on unbalanced optimal transport theory. Compared to the WMD, the WFR document distance provides a tradeoff between global transportation and local truncation, which leads to a better similarity measure for unequal semantic details. Moreover, an efficient prune strategy is particularly designed for the WFR document distance to facilitate the top-k queries among a large number of documents. Extensive experimental results show that the WFR document distance achieves higher accuracy that WMD and even its supervised variation s-WMD.

1. Introduction

Measuring the similarity between linguistic objects plays an important role in natural language processing. Word Mover's Distance (WMD) (Kusner et al., 2015) measures the Wasserstein distance of documents as bag of words distributed in word embedding space. As a mathematically solid metric, WMD comes with clear interpretation and has demonstrated great success in many applications, e.g. metric learning (Huang et al., 2016), document retrieval (Wu et al., 2018), question answering (Brokos et al., 2016) and word trans-

Table 1: Transport plan by Example 1 (WMD) and Example 2 (WFR)

	word	WMD				WFR			
		B		cost		B		cost	
		awful		amount	total	awful		amount	total
		indiv. cost	mass			indiv. cost	mass		
A	happy	1.43	1	1.43	1.43	2.04	1	2.04	2.04
C	sad	1.20	0.20	0.24	1.50	1.45	0.47	0.68	1.96
	lost	1.49	0.20	0.30		2.21	0.16	0.35	
	key	1.50	0.20	0.30		2.25	0.15	0.34	
	evening	1.56	0.20	0.31		2.42	0.12	0.30	
	restaurant	1.75	0.20	0.35		3.07	0.09	0.30	

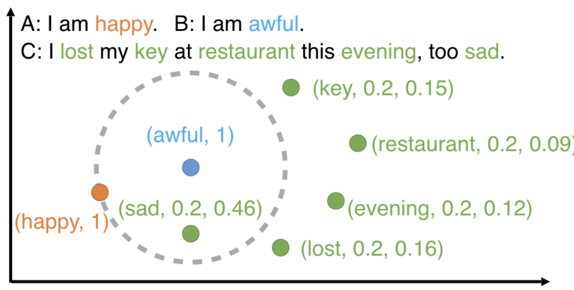

Figure 1: Illustration of transport plans by WMD (Example 1) and WFR Document Distance (Example 2). (key, 0.2, 0.15) denotes the mass of the word key is 0.2 in WMD while 0.15 in WFR.

lation (Grave et al., 2019). More concretely, in the word embedding space, WMD employs the Wasserstein metric on the space of normalized bag of words (nBOW) distribution of documents, i.e. given two documents $D_s = \{x_1^s, ..., x_m^s\}$ and $D_t = \{x_1^t, ..., x_n^t\}$ with nBOW distributions f^s and f^t, the WMD of D_s and D_t is:

$$\text{WMD}(D_s, D_t) = \min_{R \in \mathbb{R}^{m \times n}} \sum_{ij} C_{ij} R_{ij} \text{s.t.} \ \sum_j R_{ij} = f_i^s, \sum_i R_{ij} = f_j^t,$$

where $C_{ij} = \|x_i^t - x_j^s\|$ is the transport cost and $x_\cdot^\cdot$ is the word vector. With the help of optimal transport theory, WMD naturally bridges the document-level distance and the word-level dissimilarity in the embedding space.

Example 1 *Consider the three sentences in Figure 1. Indeed, sentence A has positive semantics while B and C are negative. Therefore, well-defined document distance should reveal $D_{AB} > D_{BC}$. After removing stop words, the cost to transport from B to A or C is listed in Table 1. During the transport from B to C, the mass at "awful" in B is equally allocated to the five words in C. Since four out of the five words are semantically far from "awful", the average individual cost is pulled up, which makes* $\text{WMD}(A, B) < \text{WMD}(B, C)$.

Classical optimal transport models require that every piece of mass in the source distribution is transported to an equal-weight piece of mass in the target distribution. However, this requirement appears to be too restrictive for documents classification, especially when there are words semantically far away from the motifs of the documents. Example 1 shows that WMD tends to overestimate the semantic dissimilarity when the longer document contains additional details that not involved in the shorter one. In this case, WMD may be too demanding and not an effective metric for comparing documents with rich semantic details. Especially, this situation becomes extremely severe in some advanced tasks such as text summarizing (Kedzie and McKeown, 2016) and length-varying matching (Gong et al., 2018).

To address the issues above, we introduce a robust document distance to make the justification of weights happens naturally, text-specifically without any supervision. Our new document is based the Wasserstein-Fisher-Rao (WFR) metric, a natural extension of Wasserstein metric newly developed from the theory of unbalanced optimal transport Stanislav Kondratyev and Vorotnikov (2016); Liero et al. (2016); Chizat et al. (2018a,b). Unlike traditional Wasserstein metric, WFR metric allows transport from a piece of mass to another piece with different mass by adding a penalty term accounting for the unbalanced mass. WFR document distance allows the unbalanced transport among semantic words, which naturally re-weight the transport plan based on the squared distances in word embedding space. This unique property of WFR alleviates the overestimation effects caused by WMD in a text-specific way. The following Example 2 illustrates how WFR document distance remains effective in the case where WMD fails.

Example 2 *The unbalanced transport plan from B to A or C and its cost that derives WFR document distance are listed in Table 1. As we can see, the points closer to "awful", such as "sad", are more preferable in the transport plan from B to C because its weights is higher (than 0.2). Words with less weight (less than 0.2) in transport plan like "evening" and "restaurant" are identified as outliers. This effect naturally re-weights the five words in C and the distance of "awful" to them, making the total cost to transport from B to C lower than B to A.*

The main contributions of this paper are three folds.

- We present a novel and robust document distance to address the overestimation issue of previous WMD in a text-specific and unsupervised way, especially for documents with unequal semantic details. This document distance is solidly grounded on the trade-off of global transport and local truncation in unbalanced optimal transport theory. The transport plan as a by-product of the document distance could be interpreted to identify the semantic outliers.

- We conduct extensive experiments in the tasks of varying-length matching and document classification. WFR document distance is proved to be far more robust than WMD when applied to varying-length documents. Moreover, the results of the eight document classification tasks comprehensively show the advantage of the WFR document distance over both WMD and its supervised variation s-WMD.

- We design an effective pruning strategy for fast top-k WFR document distance query. With GPU implementation, the computation efficiency is improved nearly by an order of magnitude. We also show other frameworks based on metric space (for example WME) could be benefited from WFR document distance. We believe the WFR document distance have the potential of practical use as well as further modifications.

2. Related Work

In this section, we briefly review the literature from the following three perspectives.
(a) Representation of documents. There have been many ways for documents representation. Latent Semantic Indexing Deerwester et al. (1990) and Latent Dirichlet Allocation Blei et al. (2003) are based on inferred latent variables generated by the graphical model. However, most of those models are lack of the semantic information in the word embedding space Mikolov et al. (2013). Stack denoising auto encoders Glorot et al. (2011), Doc2Vec Le and Mikolov (2014) and skip-thoughts Kiros et al. (2015) are neural network based similarities. Despite their numerical success, those models are difficult to explain, and the performance always relies on the training samples.

Recently, WMD Kusner et al. (2015) is proposed as an implicit document representation. By considering each document as a set of words in the word embedding space, it defines the minimal transportation cost as the distance between two documents. This metric is interpretable with the consideration of semantic movements. Many other metric learning models are inspired by the metric property of WMD. S-WMD Huang et al. (2016) employed the derivative of WMD to optimize the parameterized transformation in word embedding space and histogram importance vector. Word Mover's Embedding Wu et al. (2018) designed a kernel method on WMD metric space. However, those methods are still more or less suffer from the overestimation issue. They do not have the document-specific re-weight mechanism as WFR Document Distance.
(b) (Un)balanced optimal transport. Optimal transport (OT) has been one of the hottest topics of applied mathematics in the past few years. It is also closely related to some subjects in pure mathematics such as geometric analysis Ma et al. (2005); Lott and Villani (2009) and non-linear partial differential equations Froese and Oberman (2011). As the most fundamental and important object of OT, Wasserstein metric can be applied to measure the similarity of two probability distributions. The objective functions defined by this metric are usually convex, insensitive to noise, and can be effectively computed. Thus, Wasserstein metric has been deeply exploited by many researchers and has been successfully applied to machine learning Arjovsky et al. (2017), image processing Schmitz et al. (2017) and computer graphics Solomon et al. (2015).

A key condition of Wasserstein metric is that the total mass of the measures to be compared should be identical. This requirement prevents further application of Wasserstein metric as it cannot capture the features with mass difference, growth or decay. To overcome the shortage, WFR metric is proposed Stanislav Kondratyev and Vorotnikov (2016); Liero et al. (2016); Chizat et al. (2018a,b) and applied to the situations where the similarity of objects (distributions) cannot be characterized by transport alone. Thus, it is not surprising that WFR has shown great performance in many applications, e.g. image processing Chizat et al. (2018b) and tumor growth modeling Chizat and Di Marino (2017).
(c) Fast calculation of (un)balanced optimal transport. Sinkhorn algorithm Cuturi (2013) solves the entropy regularized OT problems. By reducing the entropy regularization term, the solution of each Sinkhorn iteration approximates to that of the original OT problem. A greedy coordinate descent version of Sinkhorn iteration Altschuler et al. (2017) called Greenkhorn is proposed to improve the convergence property. Recently, Sinkhorn algorithm

is applied to solve the unbalanced optimal transport problem Chizat et al. (2018b), which is the foundation of the computation of the Wasserstein-Fisher-Rao document distance

This dual lower bound is computationally cheap and could be used in prune strategy to further accelerate the KNN.

3. Backgrounds: Unbalanced Optimal Transport

Like traditional Wasserstein metric, WFR metric can be interpreted as the *square root* of the minimum cost of a transport problem. The most intuitive approach to formulate this optimization is by introducing the Benamou-Brenier formulation of optimal transport theory:

Definition 1 (WFR metric) *Given two measures μ and ν over some metric space $(X, \|\cdot\|)$ and $\eta > 0$. Then the WFR metric is defined by the following optimization problem*

$$\mathrm{WFR}_\eta(\mu,\nu) = \left(\inf_{\rho,v,\alpha} \int_{\substack{x\in\Omega \\ t\in[0,1]}} \left(\frac{1}{2}\|v\|^2 + \frac{\eta^2}{2}\alpha^2 \right) \mathrm{d}x\mathrm{d}t \right)^{\frac{1}{2}}$$

The infimum is taken over all the triplets of fields $(\rho(t,x), v(t,x), \alpha(t,x))$ satisfying the following continuity equation:

$$\partial_t \rho + \nabla \cdot (\rho v) = \rho\alpha, \quad \rho(0,\cdot) = \mu, \quad \rho(1,\cdot) = \nu.$$

The "source term" $\rho\alpha$ in the continuity equation and the corresponding penalty term $\eta^2\alpha(t,x)^2/2$ in the objective function in the formulation of WFR metric are the main differences between WFR and classical Wasserstein metric. They quantify the failure of conservation law (mass balance) in the transport plan. The parameter η controls the interpolation of the transport cost and the penalty term, which also determines the maximum distance that transport could occur. One can refer to Chizat et al. (2018a) for more details.

We note that this dynamic formulation is one of the most important motivation. For the Dirac distribution, Equation (1) reveals the *locality* (see Lemma 2), which is intrinsically different from Wasserstein metric based WMD Kusner et al. (2015).

3.1. WFR metric for Discrete distribution

Discrete measure μ over $\mathbb{R}^n$ could be considered as $\mu = \sum_i \mu_i \delta_{x_i}$, where δ_x is the Dirac function on $x \in \mathbb{R}^n$. When $\sum_i \mu_i = 1$, μ is probabilistic distribution. In the following context, we begin with the explicit formula of the transport between two Diracs and the proof is from Section 4 in Chizat et al. (2018a). This simplest situation of unbalanced optimal transport brings the clear interpretation of local truncation described in the following lemma.

Lemma 2 *Given two Diracs $\mu = h_0\delta_{x_0}$ of mass h_0 at and location x_0 and $\nu = h_1\delta_{x_1}$ of mass h_1 and location x_1, the WFR metric of two Diracs behaves in three distinct ways depending on the distance of Diracs.*

1. *Traveling Dirac: if* $|x_0 - x_1| < \pi\eta$ *then the transporting Dirac is implicitly defined by:*

$$\rho(t) = h(t)\delta_{x(t)},$$

where $h(t) = At^2 - 2Bt + h_0$, $h(t)x'(t) = \omega_0$, $A = h_1 + h_0 - 2\sqrt{\frac{h_0 h_1}{1+\tau^2}}$, $B = h_0 - \sqrt{\frac{h_0 h_1}{1+\tau^2}}$, $\omega_0 = 2\eta\tau\sqrt{\frac{h_0 h_1}{1+\tau^2}}$ *and* $\tau = \tan\left(\frac{||x_1 - x_0||}{2\eta}\right)$. *This is the unique geodesic.*

2. *(No-transport) Fisher-Rao Geodesic: if* $|x_0 - x_1| > \pi\eta$ *then*

$$\rho(t) = t^2 h_1 \delta_{x_1} + (1-t)^2 h_0 \delta_{x_0}$$

is the unique geodesic.

3. *Cut Locus: if* $|x_0 - x_1| = \pi\eta$, *there are infinite many geodesics, including traverling Dirac and Fisher-Rao Geodesic.*

To summarize, the WFR metric between two Diracs are:

$$\mathrm{WFR}_\eta(\mu, \nu) = \sqrt{2}\eta \left[h_0 + h_1 - 2\sqrt{h_0 h_1}\cos_+\left(\frac{|x_1 - x_0|}{2\eta}\right)\right]^{\frac{1}{2}},$$

where

$$\cos_+(x) = \begin{cases} \cos(x), & x \in [-\pi/2, \pi/2]; \\ 0, & x \notin [-\pi/2, \pi/2]. \end{cases}$$

It is observed that the local truncation of WFR only allows the transport of two Diracs that no further away than $\pi\eta$. This propery helps to understand the global and local trade-off in Example 2 intuitively. In general, the transport of two discrete distributions composed of multiple Diracs can be interpreted as the linear combination of point-to-point transports. Considering two distributions,

$$\mu = \sum_{i=1}^{I} \mu_i \delta_{x_i}, \quad \nu = \sum_{j=1}^{J} \nu_j \delta_{y_j}, \quad \mu_i \geq 0,\ \nu_j \geq 0,$$

The mass μ_i, ν_j are split into different pieces $\alpha_{ij} \geq 0, \beta_{ji} \geq 0$ as

$$\sum_{j=1}^{J} \alpha_{ij} = \mu_i, \quad \sum_{i=1}^{I} \beta_{ji} = \nu_j \text{ for every } i \text{ and } j$$

and assign each pair of $(\alpha_{ij}, \beta_{ji})$ to the transport between x_i and y_j. The WFR distance between μ and ν is

$$\mathrm{WFR}_\eta^2(\mu, \nu) = \min_{\alpha_{ij}, \beta_{ji}} \sum_{i,j} \mathrm{WFR}_\eta^2(\alpha_{ij}\delta_{x_i}, \beta_{ji}\delta_{y_j}) \tag{1}$$

4. Methods

4.1. WFR document distance

Under the bag of features point of view, one document should be formulated as one discrete measure $\mu = \sum_{k=1}^{K} \mu_k \delta_{x_k}$. Following the bag of words representation, a document D is considered as a multi-set with K elements $D = \{w_1, \ldots, w_K\}$ and the number of occurrence of each word $C_D = \{c_1, \ldots, c_K\}$. Each word w_i belongs to the vocabulary $\mathcal{V}$. The nBOW distribution is defined by normalizing the number of occurrences: $\mu_k = c_k / \sum_j c_j$ for $k = 1, \ldots, K$. Given a word embedding $\mathcal{X} : \mathcal{V} \mapsto \mathbb{R}^n$, each word w_k in Document D is mapped to a point in $\mathbb{R}^n$, i.e. $x_k = \mathcal{X}(w_k)$ for $k = 1, \ldots, K$. Formally, we define the WFR document distance as follows.

Definition 3 (WFR document distance) *Given a pair of documents D_1 and D_2 and a constant $\eta > 0$. Let $\mu = \sum_{i=1}^{I} \mu_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{J} \nu_j \delta_{y_j}$ be the nBOW probability distribution of D_1 and D_2 respectively. The WFR document distance between D_1 and D_2 is defined as*

$$\mathrm{Dist}(D_1, D_2) = \mathrm{WFR}_\eta(\mu, \nu).$$

It is noted that the problem of (1) is equivalent to the minimization problem in Definition 1. However, it is difficult to find a numerical method to implement (1). Theorem 4 which is more numerically friendly is derived.

Theorem 4 *Chizat et al. (2018a) Wasserstein-Fisher-Rao metric $\mathrm{WFR}_\eta(\mu, \nu)$ for two discrete measures μ, ν is the optimum of the primal problem:*

$$\mathrm{WFR}_\eta(\mu, \nu) = \inf_{R_{ij} \geq 0} J_\eta(R; \mu, \nu). \tag{2}$$

R_{ij} is the transport plan and the objective function J_η is

$$J_\eta(R; \mu, \nu) = tr(CR^\top) + \mathcal{KL}\left(R\mathbb{1} \| \mu\right) + \mathcal{KL}\left(R^\top \mathbb{1} \| \nu\right)$$

where

$$C_{ij} = -2 \log(\cos_+(|x_i - y_j|/2\eta)) \tag{3}$$

is the cost matrix, $\mathbb{1}$ indicates the column vector whose elements are all 1, and $\mathcal{KL}$ denotes the KL divergence[1] *The corresponding dual problem is*

$$\sup_{\phi_i, \psi_j} D_\eta(\phi, \psi; \mu, \nu) \quad s.t. \quad \phi_i + \psi_j \leq C_{ij} \ \text{ for any } \ i, j.$$

where the dual objection function is

$$D_\eta(\phi, \psi; \mu, \nu) = \langle 1 - e^{-\phi}, \mu \rangle + \langle 1 - e^{-\psi}, \nu \rangle, \tag{4}$$

where $\langle \cdot, \cdot \rangle$ is the inner product and ϕ, ψ are called potentials.

1. Applying this cost function in balanced OT is another modification. We did the ablation study in Section 5 to show that the KL part is also necessary.

Algorithm 1 WFRSinkhorn($\mu, \nu, C, \epsilon, n, \phi, \psi$)

Input: Discrete measure μ and ν, cost matrix C, ϵ for entropy regularization and number of iteration n
Output: Optimal transport plan R and potential ϕ, ψ.
$(b, \phi, \psi) \leftarrow (\mathbf{1}_J, \phi, \psi)$
$R_{ij} \leftarrow \exp(\frac{\phi_i + \psi_j - C_{ij}}{\epsilon})$
for $k = 1$ **to** n **do**
  $a_i \leftarrow (\mu_i / \exp(\phi_i) \sum_j R_{ij} b_j)^{1/(1+\epsilon)}$, $b_j \leftarrow (\nu_j / \exp(\psi_j) \sum_i R_{ij} a_i)^{1/(1+\epsilon)}$
  if $\|a\|$ or $\|b\|$ is too large, or k equals to n **then**
    $\phi \leftarrow \phi + \epsilon \log(a)$, $\psi \leftarrow \psi + \epsilon \log(b)$
    $R_{ij} \leftarrow \exp(\frac{\phi_i + \psi_j - C_{ij}}{\epsilon})$
    $b \leftarrow \mathbf{1}_J$
  end if
end for
Return (R, ϕ, ψ).

4.2. Numerics for WFR document distance

Sinkhorn iteration Cuturi (2013) aims at solving the family of "entropy regularized" optimal transport problems, including WFR. We use the calligraphy letter to distinguish the regularized problem from the original one.

The entropy regularized optimal transport problem is the minimization of

$$\inf_{R_{ij}>0} \mathcal{J}_{\eta,\epsilon}(R) := J_\eta(R) + \epsilon \sum_{ij} R_{ij} \log(R_{ij}), \tag{5}$$

which is strictly convex. Up to a multiplier $2\eta^2$, we have

$$\mathcal{J}_{\eta,\epsilon}(R) = \mathcal{KL}\left(R\mathbb{1}\|\mu\right) + \mathcal{KL}\left(R^\top \mathbb{1}\|\nu\right) + \epsilon\mathcal{KL}\big(R_{ij}|| \exp(-C/\epsilon)\big)$$

By convex optimization theory Rockafellar (2015), the dual problem (5) is

$$\sup_{\phi,\psi} \mathcal{D}_{\eta,\epsilon}(\phi, \psi), \tag{6}$$
$$\mathcal{D}_{\eta,\epsilon}(\phi,\psi) = \langle 1 - e^{-\phi}, \mu\rangle + \langle 1 - e^{-\psi}, \nu\rangle + \epsilon\langle 1 - e^{\frac{\phi\oplus\psi}{\epsilon}}, K_\epsilon\rangle,$$

where $K_\epsilon = e^{-C/\epsilon}$, $(\phi \oplus \psi)_{ij} = \phi_i + \psi_j$ and could be solved by alternative gradient descent Benamou et al. (2015).

Proposition 5 *Let $u = e^{\phi/\epsilon}$ and $v = e^{\psi/\epsilon}$, the Sinkhorn iteration S_ϵ for Problem (6) is*

$$\begin{aligned} u_i^{(l+1)} &= \left(\mu_i / \textstyle\sum_j e^{-C_{ij}/\epsilon} v_j^{(l)}\right)^{1/(1+\epsilon)}, \\ v_j^{(l+1)} &= \left(\nu_j / \textstyle\sum_i e^{-C_{ij}/\epsilon} u_i^{(l+1)}\right)^{1/(1+\epsilon)}. \end{aligned} \tag{7}$$

where $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

Algorithm 1 describes single Sinkhorn iteration S_ϵ that are used to calculate entropy regularized Wasserstein-Fisher-Rao metric with log-domain stabilization. It is noted that in (7), the term $e^{-C_{ij}/\epsilon}$ or u, v might be extremely small or large which could cause the numerical instability in the implementation. In the Sinkhorn algorithm, $\exp((\phi_i + \psi_j - C_{ij})/\epsilon)$ is taken as a whole for improving the numerical stability.

Algorithm 2 WFRDocDist($\mu, \nu, M, \{(\epsilon_m, n_m)\}, \eta$)

Input: Documents distribution μ and ν, number of the WFR Sinkhorn iteration M, $\{(\epsilon_m, n_m)\}_{m=1}^{M}$ for each iteration, η for WFR metric.
Output: WFR document distance
$C_{ij} \leftarrow -2\log\left(\cos_+\left(\frac{\|x_i - y_j^{(n)}\|_2}{2\eta}\right)\right)$
$(\phi, \psi) \leftarrow (\mathbf{0}, \mathbf{0})$
for m from 1 to M **do**
  $(R, u, v) \leftarrow$ WFRSinkhorn$(\mu_{D_1}, \nu_{D_2}, C, \epsilon_m, n_m, \phi, \psi)$
end for
Return $J_\eta(R; \mu, \nu)$.

To solve the original problem (2) for WFR document distance, we sequentially perform WFR Sinkhorn iterations $\{S_{\epsilon_n}\}$ on descending $\{\epsilon_n\}$ where $\epsilon_n \to 0$, and adopt the optimal ϕ, ψ for S_{ϵ_n} as the initial value for $S_{\epsilon_{n+1}}$. The precision of WFR metric is controlled by the gap between the primal and dual problem. Algorithm 2 shows how to get the WFR document distance based on Algorothm 1.

In our experiment, we use $M = 5$ WFR Sinkhorn iterations with parameter $\{(\epsilon_m, n_m) = \{(e^{-m-1}, 32m)\}\}$ for the m-th iteration. Experiments show that the mean relative error of the approximate solution is no more than 0.001 by evaluating the duality gap which achieves the desired accuracy.

4.3. Pruning strategy for top-k smallest WFR document distance query

Top-k smallest WFR document distance query is significant in applications like document retrieval. Kusner et al. (2015) proposed a pruning strategy for fast WMD-KNN classification based on the lower bound of WMD. In the case of WFR document distance, it is natural to adopt the evaluated value of the dual objective function (4) as a lower bound. With the descending of the entropy regularization's coefficient ϵ, the dual lower bound gets more and more tight.

In the top-k smallest WFR document distance query setting, the query document D_0 is formulized as $\mu_{D_0} = \sum_{i=1}^{I} \mu_i \delta_{x_i}$ and the document samples are $\{(D_n, y_n)\}_{n=1}^{N}$ where each D_n as $\nu_{D_n} = \sum_{j=1}^{J} \nu_j^{(n)} \delta_{y_j^{(n)}}$, $n = 1, \ldots, N$. Considering the task with hyper-parameter k, after each WFR Sinkhorn iteration, we sort the document samples by the value of primal objective (2) and take the maximum of WFR document distance among the first k smallest values as the threshold. Furthermore, we evaluate the dual lower bound, document samples with lower bounds that are larger than the threshold will be dropped. By this way, we only need to perform few WFR Sinkhorn iterations for most of the samples, which saves a lot of time.

For WFR document distance described in Definition 3, the number of WFR Sinkhorn iterations M and parameters $\{(\epsilon_m, n_m)\}$ for each WFR Sinkhorn iteration is fixed. Given document size L, the time complexity of the Sinkhorn iteration is $O(L^2)$ for a fixed parameter. Given the size of training samples N, the time complexity of WFR-KNN classification is bounded by $O(NL^2)$. It is noticed that this asymptotic bound cannot be further improved since the time complexity of the distance/cost matrices calculation between the evaluated sample and N labeled samples are $O(NL^2)$. Algorithm 3 describe how to accelerate the KNN calculation by pruning the lower bounds.

Algorithm 3 Top-k smallest WFR document distance query

Input: Test document D_0 and training document set $\{(D_n, y_n)_{n=1}^N\}$, number of iteration M, parameter $\{(\epsilon_m, n_m)\}_{m=1}^M$ for each WFR Sinkhorn iteration, η for WFR document distance and K for KNN.
Output: k indices of top-k smallest WFR document distance samples
for each D_n in training set **do**
  $C_{ij}^{(n)} \leftarrow -2\log\left(\frac{\cos_+\left(\|x_i - y_j^{(n)}\|_2\right)}{2\eta}\right)$, $(u^{(n)}, v^{(n)}) \leftarrow (\text{None}, \text{None})$
end for
$FilteredIndex \leftarrow [1, \ldots, N]$
for m from 1 to M **do**
  $CandidateIndex \leftarrow FilteredIndex$
  $FilteredIndex \leftarrow [\ \ ]$
  $threshold \leftarrow 0$
  for k from 1 to K **do**
    $t \leftarrow \text{WFRDocDist}(\mu_{D_0}, \nu_{D_m}, M, \{(\epsilon_m, n_m)\}, \eta)$
    if $t \geq threshold$ **then**
      $threshold \leftarrow t$
    end if
  end for
  for each $i \in CandidateIndex$ **do**
    $(R^{(i)}, u^{(i)}, v^{(i)}) \leftarrow \text{WFRSinkhorn}(\mu_{D_0}, \nu_{D_i}, C^{(i)}, \epsilon_m, n_m, u^{(i)}, v^{(i)})$
    if $D_\eta(u^{(i)}, v^{(i)}; \mu_{D_0}, \nu_{D_i}) < threshold$ **then**
      append i to $FilteredIndex$
    end if
  end for
  Sort $FilteredIndex$ by $J_\eta(R^{(i)}; \mu_{D_0}, \nu_{D_i})$ in ascending order.
end for
Return the first-K elements of $FilteredIndex$.

5. Experiment and Discussion

In this section, we demonstrate the supreme of WFR Document Distance over WMD and other WMD based metrics in two tasks. The first task directly illustrates the robustness of WFR over WMD when matching length-varying documents, which is exactly the situation that documents have unequal details. The second task examines the effectiveness of WFR Document Distance on a vast number of documents by KNN classification. The WMD is computed by the code provided by Kusner et al. (2015).

Task 1: Varying-length matching

(a) Setup. The concept-project dataset by Gong et al. (2018) is designed for length-varying document matching task. This dataset contains 537 samples. Each sample contains one short document named "concept", one long document named "project" and one human annotated binary label for whether this pair is a good match. The length of each "concept" and corresponding "project" varies a lot. The mean of the distinct words among all "concept" is 26.4, while the mean distinct words among all "project" is 556.6. The matching is binary classification. The ratio of true and false label is 56:44. For this task, we take WMD as the baseline. It has been proven Gong et al. (2018) that WMD is a stronger than the neural network methods such as doc2vec Le and Mikolov (2014). We also apply the cost function of WFR (see Equation (3)) to balanced optimal transport as an ablation study. Suggested by Gong et al. (2018), the document distance between "concept" and "project" (WMD or WFR) is used as one score of the concept-project pair for binary classification. Given the threshold, the pair with the distance smaller than the threshold is classified as the true label. The word embedding in the experiments is pretrained by fasttext Mikolov et al. (2018). We evaluate WMD and WFR Document Distance on the whole dataset. After calculating the document distance of each pair, we adjust the threshold to obtain the precision-recall curve.

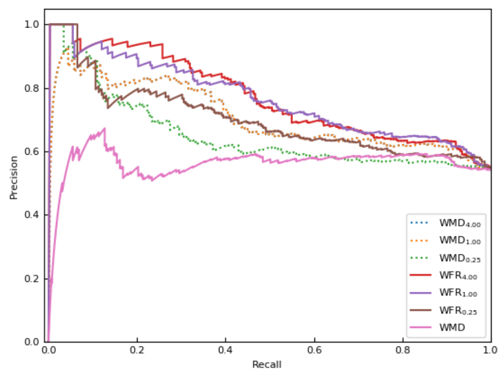

Figure 2: PR curve for the length-varying matching task

Table 2: The datasets used for evaluation and their description.

DATASET	# TRAIN	# TEST	AVG NDW	STD NDW
bbcsports	517	220	117.0	**55.0**
twitter	2175	933	9.9	5.1
recipe2	3059	1311	48.4	29.8
ohsumed	3999	5153	59.2	22.3
classic	4965	2128	38.8	27.7
reuters	5485	2189	37.1	36.6
amazon	5600	2400	45.1	**45.8**
20news	11293	7528	69.7	**70.1**

(b) Discussion. Figure 2 illustrates the precision-recall curves of WMD and WFR Document Distance whose hyper-parameter η ranges from 0.25 to 4. The curve of WMD is dominated by that of all WFR Document Distances at all recall level. At low recall level (less than 0.1, the threshold is small), the pairs with small document distances are classified to be good matches. The high precision (over 0.8) of WFR Document Distances of all hyper-parameters shows the effectiveness of our WFR Document Distance. The low precision of WMD is consistent with the observation in the Example 1 that document pair who is semantically similar may not be closed under WMD. WFR Document Distance is proved to be more reliable and robust than WMD and is not sensitive to the hyper-parameter η. ($\text{WMD}_{4.00}$ and $\text{WMD}_{1.00}$ collapsed together, which also supports η is not sensitive.) Composing ground metric of WMD with the cost function of WFR improves the performance. However, for fixed parameter η this amendment in balanced optimal transport is clearly weaker than unbalanced WFR document distance.

Task 2: KNN classification

(a) Setup. We evaluate the effectiveness of WFR Document Distance on eight document classification datasets: BBCSPORTS: BBC sports article at 2004-2005; TWITTER: sen-

Table 3: KNN classification error rate for WFR and other baselines.

DATASET	BBCSPORT	TWITTER	RECIPE	OHSUMED	CLASSIC	REUTERS	AMAZON	20NEWS
WMD	4.6 ± 0.7	28.7 ± 0.6	42.6 ± 0.3	44.5	2.8 ± 0.1	3.5	7.4 ± 0.3	28.3
S-WMD	2.1 ± 0.5	27.5 ± 0.5	39.2 ± 0.3	**34.3**	3.2 ± 0.2	**3.2**	5.8 ± 0.1	26.8
WFR	$\mathbf{1.1 \pm 0.3}$	$\mathbf{26.4 \pm 0.2}$	$\mathbf{38.9 \pm 0.1}$	41.82	$\mathbf{2.6 \pm 0.2}$	**3.2**	$\mathbf{4.8 \pm 0.2}$	**24.3**

timent classification corpus of tweets; RECIPE: recipe procedures from different origins; OHSUMED: medical abstracts from cardiovascular disease groups; CLASSIC: academical papers by different publishers; REUSTERS and 20NEWS: news articles by topics. The preprocessing procedures and the choice of word embeddings are the same as that described by Kusner et al. (2015); Huang et al. (2016). We use directly the preprocessed version of datasets from the authors. The key information of the datasets are presented in Table 2, including the number of train/test samples and the average and the standard deviation of the number of distinct words (NDW).

Besides WMD, We consider an additional supervised baseline named **Supervised Word Mover's Distance (S-WMD)**. Compared to WMD, this method employed a histogram importance vector w of vocabulary to re-weight the nBOW distribution $\tilde{f}_i = w_i f_i / \sum_j w_j f_j$, and a linear transformation $A : x_i \mapsto Ax_i$ to modify the distances in the word embedding space. The parameters are trained by gradient descent of the loss defined by Neighborhood Components Analysis (NCA). Other traditional document representation or similarity baselines are proved to be significantly weaker than WMD and SWMD Kusner et al. (2015); Huang et al. (2016). So they are not included. Throughout the experiments, we optimize over the neighborhood size ($k \in \{1, \ldots, 19\}$) in KNN and the only hyper-parameter ($\eta \in \{1, 1/2, 1/3, 1/4\}$) by 5-fold cross-validation. We obtain the original code from the authors and re-conduct the evaluation process. For datasets without predefined train/test splits (bbcsport, twitter, recipe, classic, amazon), we report the mean and standard deviation of the performance over five random 70/30 train/test splits.

(b) Discussion. In the first three rows of Table 3 we output the results from three different document distances and eight datasets. Firstly, we compare the performance between WFR with WMD. As presented, WFR Document Distance has less KNN classification error rate at all datasets. Furthermore, for the datasets with large standard deviation of NDW (exceeds 40), i.e. dataset BBCSPORTS, AMAZON and 20NEWS, WFR outperforms the document distance with a clear margin. For those datasets with less standard deviations of NDW, the reduction of the KNN classification error is not that significant. Secondly, we compare the performance between WFR with S-WMD. WFR successfully outperforms S-WMD in six out of eight datasets even though S-WMD has more supervised parameters. The successful of WFR over S-WMD since a more effective way to re-weight the transport plan is automatically captured by WFR, rather than text-independent global re-weighting in S-WMD. We notice that S-WMD only outperforms WFR and WMD at OHSUMED dataset. The medical term for cardiovascular disease in the OHSUMED dataset may not have proper word vector. The text-independent deficiency of the word embedding might be relieved by supervision in S-WMD.

(c) WFR Document Distance for Other Frameworks. Word Mover's Embedding (WME Wu et al. (2018)) framework is proposed to abstract the document space of Word Mover's Distance (or other metric spaces). This framework realized fast estimation of WMD

Table 4: KNN classification error rate for WME+WFR and WME+WFR.

DATASET	BBCSPORT	TWITTER	RECIPE	OHSUMED	CLASSIC	REUTERS	AMAZON	20NEWS
WME(512)+WMD	3.5 ± 0.7	26.8 ± 2.3	48.0 ± 0.6	42.1	4.8 ± 0.3	4.0	**7.4 ± 0.4**	30.7
WME(512)+WFR	**2.7 ± 1.0**	**26.0 ± 1.9**	**43.3 ± 1.2**	**37.2**	4.8 ± 0.3	4.0	**7.4 ± 0.4**	30.7
WME(4096)+WMD	2.0 ± 1.0	25.9 ± 2.3	40.2 ± 0.6	36.5	3.0 ± 0.3	2.8	5.7 ± 0.4	22.1
WME(4096)+WFR	**1.8 ± 1.0**	**25.3 ± 1.6**	**40.1 ± 0.6**	**34.4**	**2.9 ± 0.3**	**2.3**	**5.5 ± 0.5**	**21.5**

Table 5: KNN prune efficiency and GPU acceleration ratio for eight datasets

DATASET	Prune			Acc.
	1st	2nd	3rd	Ratio
BBCSPORTS	89.6%	6.2%	4.1%	3.9
TWITTER	2.1%	1.0%	0.9%	35.0
RECIPE2	46.3%	1.9%	0.8%	7.3
OHSUMED	31.1%	0.8%	0.5%	8.9
CLASSIC	33.8%	1.2%	0.5%	9.0
REUTERS	3.2%	0.6%	0.4%	11.8
AMAZON	1.7%	0.5%	0.4%	9.9
20NEWS	40.0%	0.6%	0.2%	6.7

by Monte Carlo's method. We found that replacing the WMD in WME framework with WFR document distances effectively improves the original results. Here we demonstrate the detailed results for WFR+WME and WMD+WME (WFR document distance and word mover's distance within WME framework). We present WME with two sizes of Monte Carlo samples, i.e. 512 and 4096. We didn't exact recover the original results in Wu et al. (2018) since we didn't know exact value of hyperparameters D_{max}, γ and R that are used for each dataset. By similar parameter selection process, we produce the compatible results (WME(512)+WMD is close to WME(SR) in Wu et al. (2018) and WME(4096)+WMD is close to WME(LR)). For WME+WFR, we take an additional cross-validation process to select the hyperparameter η of WFR. The differences of WME and WFR are compared under the same MC sample condition. In most case, We could see that WME+WMD is weaker than WME+WFR for both 512 and 4096 MC samples. For some datasets (TWITTER, OHSUMED and CLASSIC), WME(512)+WFR is really close to the WME(4096)+WMD. This results support that WFR document distance is better than word mover's distance.

5.1. Pruning Efficiency and Time Cost

The pruning strategy for top-k smallest WFR document distance query and GPU parallelism is important to constrain the computation cost of KNN in an affordable range. Table 5 demonstrates the effect of prune strategy and GPU parallelism.

The columns in Table 5 named by Prune shows the average percent of samples left after m-th round of prune. For BBCSPORTS dataset, since the training set has only 517 samples, 3.87% of the training set contains 20 samples, which is the minimal number required for KNN classifier when $K = 20$. For other larger datasets, we noticed that after 2 rounds, more than 98% of the training samples are pruned. For all datasets, one could examine that after 3 rounds, the number of left samples is about 20, which is suitable for the following KNN classification. With this pruning strategy, most of the computing cost is at the 1st Sinkhorn iteration, which is of time complexity $O(NL^2)$. In other words, one could improve the final precision for top-k smallest WFR document distance with merely little cost.

Figure 3 shows the averaging time of one KNN classification on eight datasets. The value is scaled by the minimal time cost (TWITTER dataset by GPU). The column in

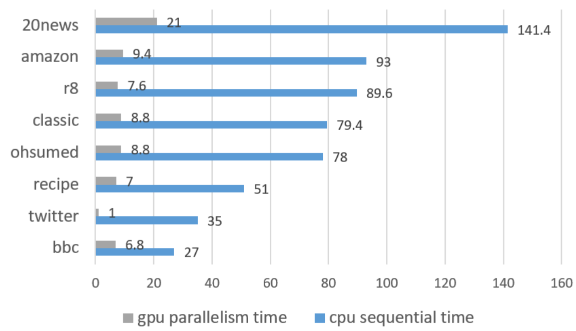

Figure 3: Scaled time cost for one KNN classification (K=20)

Table 5 named by GPU Acc. Ratio denotes the acceleration ratio. For example, 3.9 for bbcsports means that one CPU (Core i7-7700HQ) computation costs 3.9 times as one GPU (GTX-1080Ti). For TWITTER, this dataset is too small so that all the data could be placed into the visual memory of GPU at single batch, which allows extremely high parallelism. Discard the highest and lowest value of the acceleration ratio, we observe that the GPU parallelism provides about 8.9 times acceleration.

Another concerning about the computation time is the difference between WFR and WMD by Sinkhorn iteration. We evaluate conduct 1000 pairs of documents (point clouds) with number of distinct words varying from 10 to 1000. Experiments are conducted by MATLAB with fixed iteration parameters. WFR takes 42.3 seconds in total while WMD takes 39.7 seconds. We think about additional 5% time cost is worthwhile.

6. Conclusion

In this paper, we present WFR document distance to address the overestimation issue of previous optimal transport. This document distance is robust to unequal semantic details by taking the advantage of the trade-off of global transportation and local truncation by unbalanced optimal transport theory. As a result, WFR document distance benefits from the dissimilarity of word-level while achieves automatically text-specific and unsupervised re-weighted transport plan. This makes it outperforms WMD and supervised s-WMD and could be seamlessly adapted to other frameworks. Extensive experiments confirm the effectiveness and efficiency of the new proposed document distance.

7. Acknowledgment

Yong Zhang were supported by NSFC (91646202), National Key R&D Program of China (2018YFB1404401, 2018YFB1402701). Hao Wu was supported by NSFC (11871297) and Tsinghua University Initiative Scientific Research Program.

References

Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *NIPS*, pages 1961–1971, 2017.

Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM J. Sci. Comput.*, 37(2), 2015.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

Georgios-Ioannis Brokos, Prodromos Malakasiotis, and Ion Androutsopoulos. Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering. In *BioNLP@ACL*, pages 114–118, 2016.

Lénaïc Chizat and Simone Di Marino. A tumor growth model of hele-shaw type as a gradient flow. *arXiv preprint arXiv:1712.06124*, 2017.

Lenaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and fisher-rao metrics. *Found. Comut. Math.*, 18(1):1–44, 2018a.

Lenaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Math. Comput.*, 87(314):2563–2609, 2018b.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, pages 2292–2300, 2013.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

Brittany D. Froese and Adam M. Oberman. Convergent finite difference solvers for viscosity solutions of the elliptic monge-ampère equation in dimensions two and higher. *SIAM J. Numer. Anal.*, 49(4):1692–1714, 2011.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, pages 513–520, 2011.

Hongyu Gong, Tarek Sakakini, Suma Bhat, and Jinjun Xiong. Document similarity for texts of varying lengths via hidden topics. In *ACL*, pages 2341–2351, 2018.

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *AISTATS*, pages 1880–1890, 2019.

Gao Huang, Chuan Guo, Matt J. Kusner, Yu Sun, Fei Sha, and Kilian Q. Weinberger. Supervised word mover's distance. In *NIPS*, pages 4862–4870, 2016.

Chris Kedzie and Kathleen R. McKeown. Extractive and abstractive event summarization over streaming web text. In *IJCAI*, pages 4002–4003, 2016.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *NIPS*, pages 3294–3302, 2015.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In *ICML*, pages 957–966, 2015.

Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, pages 1188–1196, 2014.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal transport in competition with reaction: The hellinger-kantorovich distance and geodesic curves. *SIAM J. Math. Analysis*, 48(4):2869–2911, 2016.

John Lott and Cédric Villani. Ricci curvature for metric-measure spaces via optimal transport. *Ann. Math.*, 169(3):903–991, 2009.

Xinan Ma, Neil S. Trudinger, and Xu-Jia Wang. Regularity of potential functions of the optimal transportation problem. *Arch. Ration. Mech. Anal.*, 177:151–183, 01 2005.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *LREC*, 2018.

Ralph Tyrell Rockafellar. Convex analysis. 2015.

Morgan A. Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Maurice Ngolè Mboula, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised non-linear dictionary learning. *CoRR*, abs/1708.01955, 2017.

Justin Solomon, Fernando de Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas J. Guibas. Convolutional wasserstein distances: efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4):66:1–66:11, 2015.

Léonard Monsaingeon Stanislav Kondratyev and Dmitry Vorotnikov. A new optimal transport distance on the space of finite radon measures. *Adv. Differ. Equ.*, 21:1117–1164, 2016.

Lingfei Wu, Ian En-Hsu Yen, Kun Xu, Fangli Xu, Avinash Balakrishnan, Pin-Yu Chen, Pradeep Ravikumar, and Michael J. Witbrock. Word mover's embedding: From word2vec to document embedding. In *EMNLP*, pages 4524–4534, 2018.