# Supervised Tree-Wasserstein Distance

*Yuki Takezawa[1 2], Ryoma Sato[1 2], Makoto Yamada[1 2]*

*[1]Kyoto University, [2]RIKEN AIP*

KYOTO UNIVERSITY

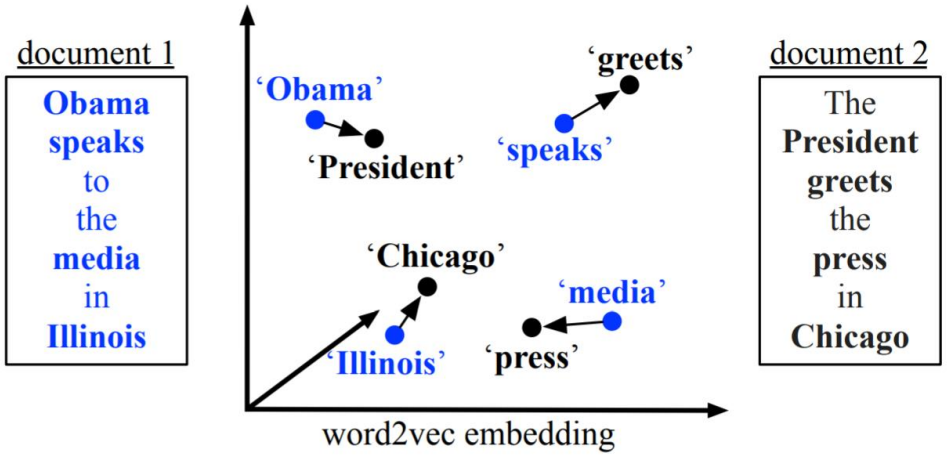
# Motivation

## The computational cost of the Wasserstein distance.

**Wasserstein distance**

$$W_d(\mu_i, \mu_j) = \inf_{\gamma \in \Pi(\mu_i, \mu_j)} \int_{\Omega \times \Omega} d(x, y)\gamma(dx, dy)$$

- Linear Programming : cubic time.
- Sinkhorn Algorithm : quadratic time.

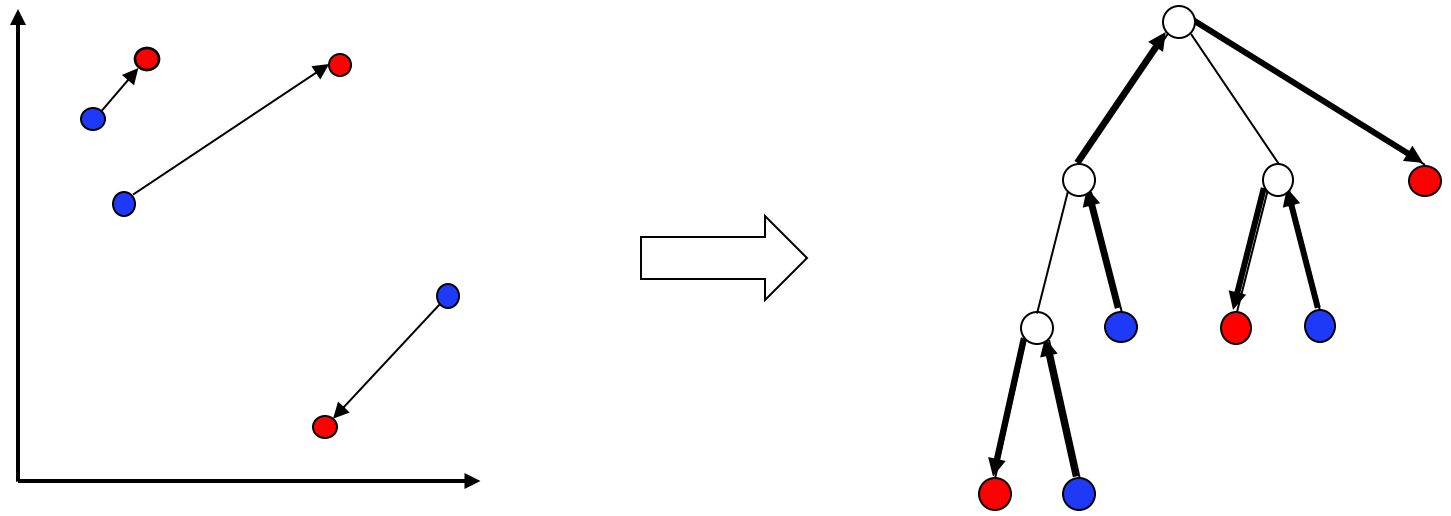



Word Mover's Distance [Kusner 15]

# Motivation
## Tree-Wasserstein Distance

The Wasserstein distance on a tree metric.

- The closed-form solution, which can be computed in linear time.



Quadtree [Indyk 03], clustering based method [Le 19]

- Unsupervised methods

# Motivation
## Supervised Tree-Wasserstein Distance

The distance between documents depends on the task.

- Topic, Author etc.
- Supervised WMD [Huang 16]
  - High computational cost.

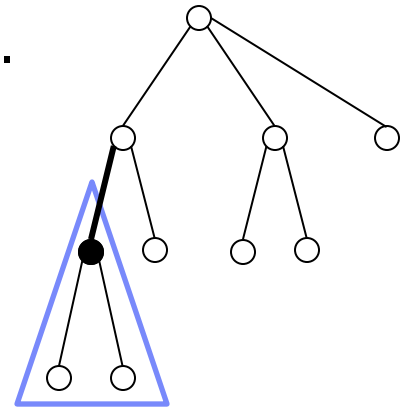
We propose the **Supervised Tree-Wasserstein (STW) Distance.**

- Task-specific distance.
- Fast computation for the Wasserstein distance.
  - Tree-Wasserstein distance
  - GPU suitable. (e.g., batch processing)

# Proposed Method
## Soft Tree-Wasserstein Distance

**Tree-Wasserstein Distance**

- $\Gamma(v)$ : the set of nodes contained in the subtree rooted at $v$.



$$W_{d_T}(\mu_i, \mu_j) = \sum_{v \in V} w_v \left| \mu_i(\Gamma(v)) - \mu_j(\Gamma(v)) \right|$$

$$= \sum_{v \in V} w_v \left| \sum_{x \in \Gamma(v)} \mu_i(x) - \mu_j(x) \right|$$

**Soft Tree-Wasserstein Distance**

- $P_{\mathrm{sub}}(x|v)$ : the probability that node $x$ is contained in the subtree rooted at $v$.

$$\rightarrow \sum_{v \in V} w_v \left| \sum_{x \in V} P_{\mathrm{sub}}(x|v)(\mu_i(x) - \mu_j(x)) \right|$$

# Proposed Method

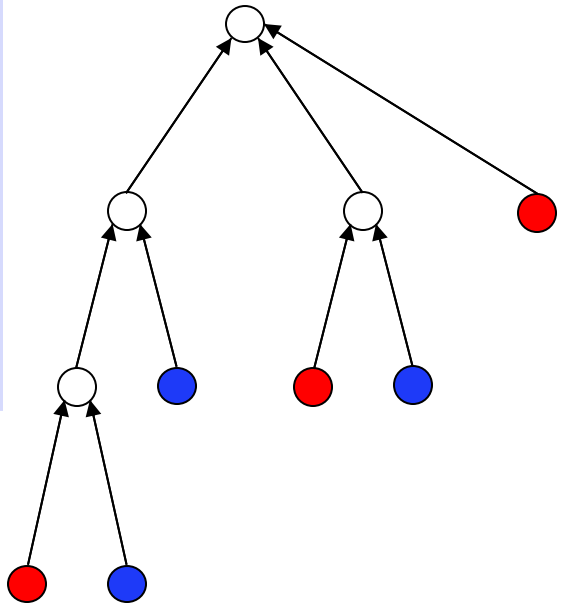
## How to formulate $P_{\mathrm{sub}}(x|v)$ ?

**Theorem 1 : conditions of an adjacency matrix**

Let $\mathbf{D}_{\mathrm{par}} \in \{0,1\}^{|V|\times|V|}$ be an adjacency matrix of the directed graph $G$.
If $\mathbf{D}_{\mathrm{par}}$ satisfies the followings:

- $\mathbf{D}_{\mathrm{par}}$ is a strictly upper triangular matrix.
- $\mathbf{D}_{\mathrm{par}}^{\top}\mathbf{1} = (0,1,\ldots,1)^{\top}$ .

Then $G$ is a directed tree.



Probability that $v_i$ is a parent of $v_j$.

- $\left[\mathbf{D}_{\mathrm{par}}\right]_{i,j} \in [0,1]$

Probability that $v_i$ is contained in the subtree rooted at $v_j$.

- $\left[\sum_{k=1}^{\infty}\mathbf{D}_{\mathrm{par}}^{k}\right]_{i,j} = \left[\left(\mathbf{I}-\mathbf{D}_{\mathrm{par}}\right)^{-1}\right]_{i,j}$

# Proposed Method
## Supervised Tree-Wasserstein (STW) Distance

**Soft Tree-Wasserstein Distance**

- Differentiable w.r.t. parent-child relationships. (i.e., $\mathbf{D}_{\mathrm{par}}$)

$$W_{d_T}^{\mathrm{soft}}(\mu_i, \mu_j) = \sum_{v \in \boldsymbol{V}} w_v \left| \sum_{x \in \boldsymbol{V}} P_{\mathrm{sub}}(x|v) \left( \mu_i(v) - \mu_j(v) \right) \right|_{\alpha}$$

$$= \left\| \boldsymbol{w}_v \circ \left( \mathbf{I} - \mathbf{D}_{\mathrm{par}} \right)^{-1} \left( \mathbf{a}_i - \mathbf{a}_j \right) \right\|_{\alpha}$$

**Objective function:**

$$\underset{\mathbf{D}_{\mathrm{par}}}{\mathrm{minimize}} \sum_{i,j:\text{same label}} W_{d_T}^{\mathrm{soft}}(\mu_i, \mu_j) - \sum_{i,j:\text{different label}} \min\left\{ W_{d_T}^{\mathrm{soft}}(\mu_i, \mu_j), m \right\}$$

$$\text{s.t. } \mathbf{D}_{\mathrm{par}} \text{ satisfies the conditions of Th. 1.}$$

# Experiment
## Document Classification Task

Table 1: The $k$NN test error for real datasets. The most accurate method based on the tree-Wasserstein distance is shown in blue.

| | TWITTER | AMAZON | CLASSIC | BBCSPORT | OHSUMED | REUTERS |
|---|---|---|---|---|---|---|
| WMD | 28.7 ± 0.6 | 7.4 ± 0.3 | **2.8 ± 0.1** | 4.6 ± 0.7 | 44.5 | 3.5 |
| S-WMD | **27.5 ± 0.5** | **5.8 ± 0.1** | 3.2 ± 0.2 | **2.1 ± 0.5** | **34.3** | **3.2** |
| QUADTREE | 30.4 ± 0.8 | 10.7 ± 0.3 | 4.1 ± 0.4 | 4.5 ± 0.5 | 44.0 | 5.2 |
| FLOWTREE | 29.8 ± 0.9 | 9.9 ± 0.3 | 5.6 ± 0.6 | 4.7 ± 1.1 | 44.4 | 4.7 |
| TSW-1 | 30.2 ± 1.3 | 14.5 ± 0.6 | 5.5 ± 0.5 | 12.4 ± 1.9 | 58.4 | 7.5 |
| TSW-5 | 29.5 ± 1.1 | 9.2 ± 0.1 | 4.1 ± 0.4 | 11.9 ± 1.3 | 51.7 | 5.8 |
| TSW-10 | 29.3 ± 1.0 | 8.9 ± 0.5 | 4.1 ± 0.6 | 11.4 ± 0.9 | 51.1 | 5.4 |
| STW | 28.9 ± 0.7 | 10.1 ± 0.7 | 4.4 ± 0.7 | 3.4 ± 0.8 | 40.2 | 4.4 |

Tree-Wasserstein distance

# Experiment
## Time Consumption

Tree-Wasserstein distance

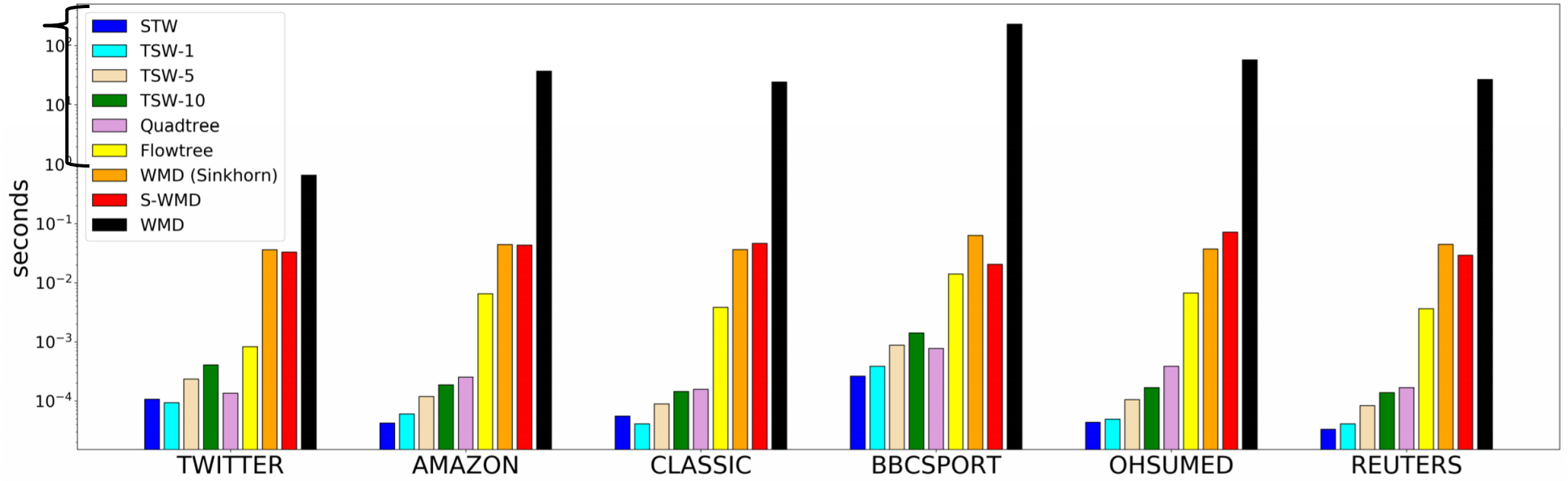



*Figure 3.* Average time consumption for comparing 500 documents with one document. For the STW distance and the TSW distance, the batch size is set to the number of documents contained in the training dataset. For WMD (Sinkhorn) and S-WMD, the batch size is set to 500 due to the memory size limitations. To obtain the average time consumption, we sample 100 documents as queries and measure the time consumption.