Fixed Support Tree-Sliced Wasserstein Barycenter

Yuki Takezawa[1 2], Ryoma Sato[1 2], Zornitsa Kozareva[3], Sujith Ravi[4], Makoto Yamada[1 2]

[1]Kyoto University, [2]RIKEN AIP, [3]Facebook AI Research, [4]SliceX AI

1. Background

Wasserstein Distance

- Powerful tool to measure the distance between distributions.
- High computational cost. (e.g., Sinkhorn algorithm requires $O(M^2)$)

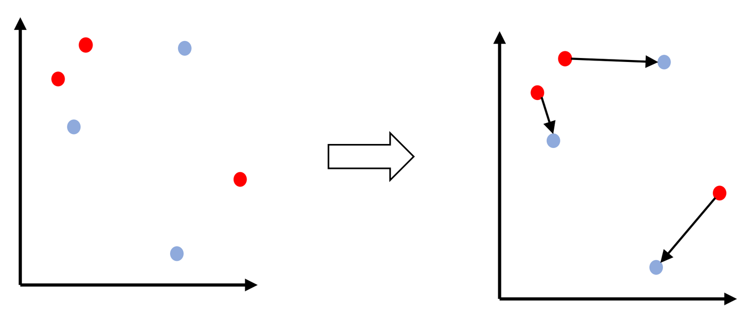


Fig.1

Wasserstein Barycenter

- $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_N \mid \mathbf{a}_i \in \mathbf{R}^M\}$: a set of probability distributions.
- $\boldsymbol{A}$: simplex.

$$\mathrm{argmin}_{\mathbf{a} \in A} \frac{1}{N} \sum_{i=1}^{N} W_d(\mathbf{a}_i, \mathbf{a})$$

Application:

- Topic Modeling [Xu+ 18]
- Generative Model [Simon+ 20]

Fig.4: [Simon+ 20]

Algorithm: Iterative Bregman Projection (IBP) [Benamou+15]

- **Time Complexity:** $O(NM^2)$ (N : #samples, M: #supports)
- The reason of this high computational cost is the computation of the Wasserstein distance itself requires $O(M^2)$.

Sliced-Wasserstein Distance

- Closed-form solution.

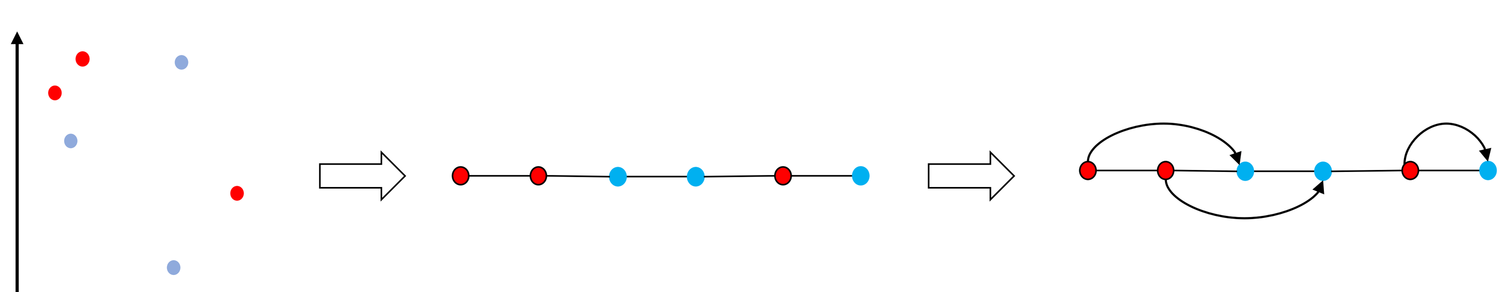


Fig.2

Tree-Sliced Wasserstein Distance

- Closed-form solution, which in linear time.
- Generalization of the sliced-Wasserstein distance.

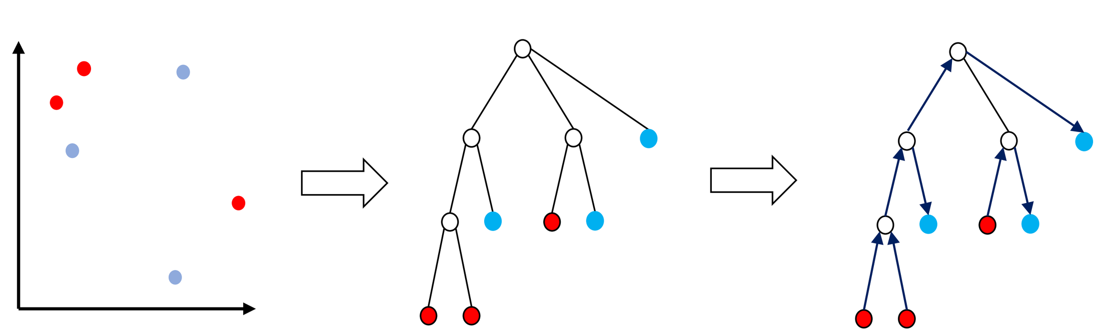


Fig.3

2. Contribution Summary

- To compute the Wasserstein barycenter fast, we propose the **Fixed Support Tree-Sliced Wasserstein Barycenter (FS-TSWB)**, the Wasserstein barycenter on the tree metric.

Proposed Algorithm:

	Time Complexity
FS-WB (IBP [Benamou+15])	$O(NM^2)$
FS-TSWB (PSD)	$O(T(\log M + N + D)M)$
FS-TSWB (FastPSD)	$O(T(\log M + \log N + D)M)$

Experimental Results:

- The FS-TSWB can be solved **2-order magnitude faster** than the FS-WB.
- The FS-TSWB can better approximate the FS-WB than the Fixed Support Sliced-Wasserstein Barycenter (FS-SWB)

N: the number of samples, M: the number of supports, D: the depth of a tree.

3. FS-TSWB

Matrix-Form Formulation of Tree Wasserstein Distance [Takezawa+ 21]

$$W_{d_T}(\mathbf{a}_i, \mathbf{a}_j) = \|\mathbf{B}(\mathbf{a}_i - \mathbf{a}_j)\|_1$$

Problem Formulation of FS-TSWB:

- $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_N \mid \mathbf{a}_i \in \mathbf{R}^M\}$: a set of probability distributions.
- $\boldsymbol{A}$: simplex.

$$\mathrm{argmin}_{\mathbf{a} \in A} \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{B}(\mathbf{a} - \mathbf{a}_i)\|_1$$

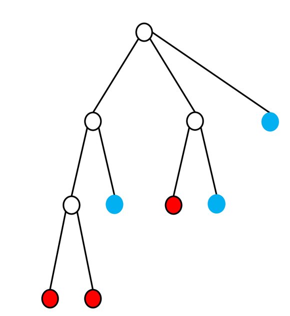


Fig.4

- Convex optimization.
- **Projected Subgradient Descent (PSD)** is applicable.

4. Algorithm: PSD

1. Compute subgradient.

$$\mathbf{g}^{(k)} = \frac{1}{N} \mathbf{B}^\top \left(\sum_i \mathrm{sign}(\mathbf{B}\mathbf{a}^{(k)} - \mathbf{B}\mathbf{a}_i) \right) \qquad O((N+D)M)$$

2. Update barycenter.

$$\mathbf{a}^{(k+1)} = \mathrm{argmin}_{\mathbf{a} \in A} \left\| \mathbf{a} - \left(\mathbf{a}^{(k)} - \gamma\, \mathbf{g}^{(k)}\right) \right\|^2 \qquad O(M \log(M))$$

3. Compute loss. $\frac{1}{N} \sum_{i=1}^{N} \|\mathbf{B}\mathbf{a}^{(k+1)} - \mathbf{B}\mathbf{a}_i\|_1 \qquad O(NM)$

Time Complexity: $O(T(\log M + N + D)M)$

- The PSD is faster than the IBP.

N: the number of samples, M: the number of supports, D: the depth of a tree.

5. Algorithm: FastPSD

Fast Computation of Subgradient:

$$\mathbf{g}^{(k)} = \frac{1}{N} \mathbf{B}^\top \left(\overbrace{\sum_i \mathrm{sign}(\underbrace{\mathbf{B}\mathbf{a}^{(k)}}_{\mathbf{b}^{(k)}} - \underbrace{\mathbf{B}\mathbf{a}_i}_{\mathbf{b}_i})}^{\boldsymbol{z}^{(k)} = \sum_{i=1}^{N} \mathrm{sign}(\mathbf{b}^{(k)} - \mathbf{b}_i)} \right)$$

$$\cancel{O((N+D)M)}$$

$$O((\log(N) + D)M)$$

Naively, to compute $[\boldsymbol{z}^{(k)}]_j = \sum_{i=1}^{N} \mathrm{sign}\left([\mathbf{b}^{(k)}]_j - [\mathbf{b}_i]_j\right)$ requires $O(N)$.

Proposed Algorithm: $O(\log N)$

These does not depend on iteration k.

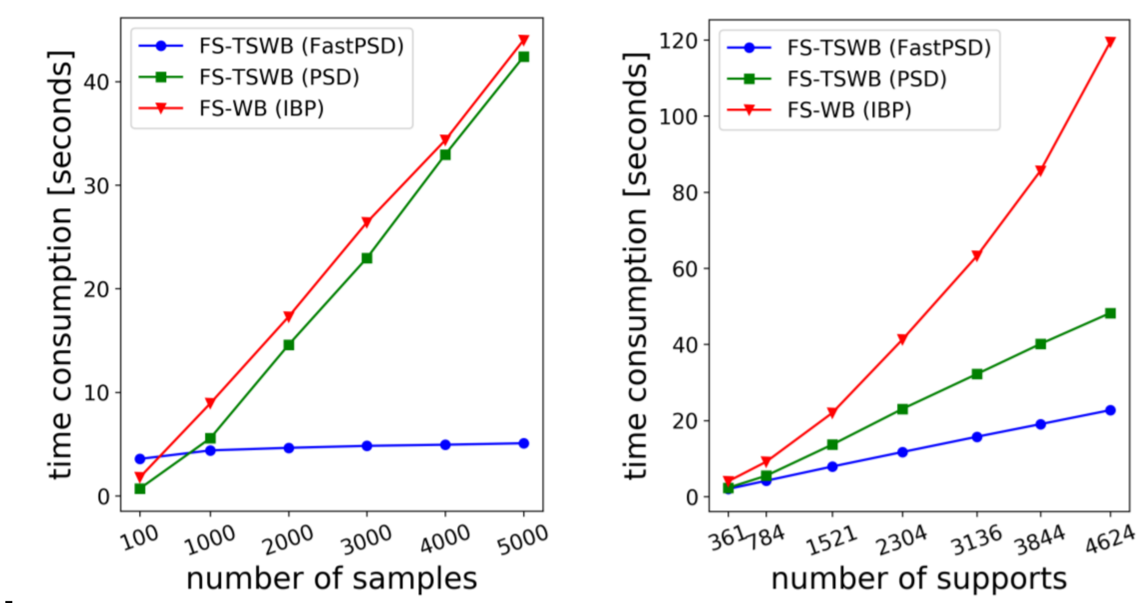
1. $[\mathbf{b}_1]_j, [\mathbf{b}_2]_j, [\mathbf{b}_3]_j, [\mathbf{b}_4]_j, [\mathbf{b}_5]_j$
2. $[\mathbf{b}_{\sigma(1)}]_j, [\mathbf{b}_{\sigma(2)}]_j, [\mathbf{b}_{\sigma(3)}]_j, [\mathbf{b}_{\sigma(4)}]_j, [\mathbf{b}_{\sigma(5)}]_j$ (sort in ascending order)
3. $\underbrace{[\mathbf{b}_{\sigma(1)}]_j, [\mathbf{b}_{\sigma(2)}]_j, [\mathbf{b}_{\sigma(3)}]_j}_{\mathrm{sign}\left([\mathbf{b}^{(k)}]_j - [\mathbf{b}_i]_j\right) = 1}, [\mathbf{b}^{(k)}]_j, \underbrace{[\mathbf{b}_{\sigma(4)}]_j, [\mathbf{b}_{\sigma(5)}]_j}_{\mathrm{sign}\left([\mathbf{b}^{(k)}]_j - [\mathbf{b}_i]_j\right) = -1}$ (binary search $O(\log N)$)

6. Experimental Results

Time Consumption: Table 1: Time consumption [seconds].

		MNIST	AMAZON	AGNews
	FS-WB	64.4	2129.6	10811.7
Propose method	FS-TSWB ($T=1$)	5.2	62.4	86.1
	FS-TSWB ($T=5$)	25.1	330.7	449.4
	FS-TSWB ($T=10$)	51.1	653.9	899.9
	FS-TSWB ($T=15$)	78.9	969.5	1346.6
	FS-TSWB ($T=20$)	111.9	1287.3	1788.4
	FS-TSWB ($T=25$)	142.6	1610.3	2236.4

$\times$ **125**



Visualization: