Supplementary Material for Multilevel Clustering via Wasserstein Means

Nhat Ho [1] **XuanLong Nguyen** [1] **Mikhail Yurochkin** [1] **Hung Hai Bui** [2] **Viet Huynh** [3] **Dinh Phung** [3]

Appendix A

In this appendix, we collect relevant information on the Wasserstein metric and Wasserstein barycenter problem, which were introduced in Section 2 in the paper. For any Borel map $g : \Theta \to \Theta$ and probability measure G on Θ, the push-forward measure of G through g, denoted by $g\#G$, is defined by the condition that $\int\limits_{\Theta} f(y)d(g\#G)(y) = \int\limits_{\Theta} f(g(x))dG(x)$ for every continuous bounded function f on Θ.

Wasserstein metric When $G = \sum\limits_{i=1}^{k} p_i\delta_{\theta_i}$ and $G' = \sum\limits_{i=1}^{k'} p'_i\delta_{\theta'_i}$ are discrete measures with finite support, i.e., k and k' are finite, the Wasserstein distance of order r between G and G' can be represented as

$$W_r^r(G, G') = \min_{T\in\Pi(G,G')} \langle T, M_{G,G'}\rangle \tag{1}$$

where we have

$$\Pi(G, G') = \left\{T \in \mathbb{R}_+^{k\times k'} : T\mathbb{1}_{k'} = \boldsymbol{p},\ T\mathbb{1}_k = \boldsymbol{p'}\right\}$$

such that $\boldsymbol{p} = (p_1, \ldots, p_k)^T$ and $\boldsymbol{p'} = (p'_1, \ldots, p'_{k'})^T$, $M_{G,G'} = \left\{\|\theta_i - \theta'_j\|\right\}_{i,j} \in \mathbb{R}_+^{k\times k'}$ is the cost matrix, i.e. matrix of pairwise distances of elements between G and G', and $\langle A, B\rangle = \text{tr}(A^T B)$ is the Frobenius dot-product of matrices. The optimal $T \in \Pi(G, G')$ in optimization problem (1) is called the optimal coupling of G and G', representing the **optimal transport** between these two measures. When $k = k'$, the complexity of best algorithms for finding the optimal transport is $O(k^3 \log k)$. Currently, (Cuturi, 2013) proposed a regularized version of (1) based on Sinkhorn distance where the complexity of finding an approximation of the optimal transport is $O(k^2)$. Due to its favorably fast computation, throughout the paper we shall utilize Cuturi's algorithm to compute the Wasserstein distance between G and $G\prime$ as well as their optimal transport in (1).

Wasserstein barycenter As introduced in Section 2.2 in the paper, for any probability measures $P_1, P_2, \ldots, P_N \in \mathcal{P}_2(\Theta)$, their Wasserstein barycenter $\overline{P}_{N,\lambda}$ is such that

$$\overline{P}_{N,\lambda} = \mathop{\arg\min}_{P\in\mathcal{P}_2(\Theta)} \sum_{i=1}^{N} \lambda_i W_2^2(P, P_i)$$

where $\lambda \in \Delta_N$ denote weights associated with $P_1, \ldots, P_N$. According to (Agueh and Carlier, 2011), $P_{N,\lambda}$ can be obtained as a solution to so-called multi-marginal optimal transporation problem. In fact, if we denote T_k^1 as the measure preseving map from P_1 to P_k, i.e., $P_k = T_k^1\#P_1$, for any $1 \leq k \leq N$, then

$$\overline{P}_{N,\lambda} = \left(\sum_{k=1}^{N} \lambda_k T_k^1\right)\#P_1.$$

Unfortunately, the forms of the maps T_k^1 are analytically intractable, especially if no special constraints on $P_1, \ldots, P_N$ are imposed.

Recently, (Anderes et al., 2015) studied the Wasserstein barycenters $\overline{P}_{N,\lambda}$ when $P_1, P_2, \ldots, P_N$ are finite discrete measures and $\lambda = \left(1/N, \ldots, 1/N\right)$. They demonstrate the following sharp result (cf. Theorem 2 in (Anderes et al., 2015)) regarding the number of atoms of $\overline{P}_{N,\lambda}$

Theorem A.1. *There exists a Wasserstein barycenter $\overline{P}_{N,\lambda}$ such that supp$(\overline{P}_{N,\lambda}) \leq \sum\limits_{i=1}^{N} s_i - N + 1$.*

Therefore, when $P_1, \ldots, P_N$ are indeed finite discrete measures and the weights are uniform, the problem of finding Wasserstein barycenter $\overline{P}_{N,\lambda}$ over the (computationally large) space $\mathcal{P}_2(\Theta)$ is reduced to a search over a smaller space $\mathcal{O}_l(\Theta)$ where $l = \sum\limits_{i=1}^{N} s_i - N + 1$.

Appendix B

In this appendix, we provide proofs for the remaining results in the paper. We start by giving a proof for the tran-

[1] Department of Statistics, University of Michigan, Ann Arbor, USA. [2] Adobe Research. [3] Center for Pattern Recognition and Data Analytics (PRaDA), Deakin University, Australia. Correspondence to: Nhat Ho <minhnhat@umich.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

sition from multilevel Wasserstein means objective function to objective function (4) in Section 3.1 in the paper. All the notations in this appendix are similar to those in the main text. For each closed subset $\mathcal{S} \subset \mathcal{P}_2(\Theta)$, denote the Voronoi region generated by $\mathcal{S}$ on the space $\mathcal{P}_2(\Theta)$ by the collection of subsets $\{V_P\}_{P \in \mathcal{S}}$, where $V_P := \{Q \in \mathcal{P}_2(\Theta) : W_2^2(Q,P) = \min\limits_{G \in \mathcal{S}} W_2^2(Q,G)\}$. We define the projection mapping $\pi_{\mathcal{S}}$ as: $\pi_{\mathcal{S}} : \mathcal{P}_2(\Theta) \to \mathcal{S}$ where $\pi_{\mathcal{S}}(Q) = P$ as $Q \in V_P$. Note that, for any $P_1, P_2 \in \mathcal{S}$ such that V_{P_1} and V_{P_2} share the boundary, the values of $\pi_{\mathcal{S}}$ at the elements in that boundary can be chosen to be either P_1 or P_2. Now, we start with the following useful lemmas.

Lemma B.1. *For any closed subset $\mathcal{S}$ on $\mathcal{P}_2(\Theta)$, if $\mathcal{Q} \in \mathcal{P}_2(\mathcal{P}_2(\Theta))$, then $E_{X\sim\mathcal{Q}}(d^2_{W_2}(X,\mathcal{S})) = W_2^2(\mathcal{Q}, \pi_{\mathcal{S}}\#\mathcal{Q})$ where $d^2_{W_2}(X,\mathcal{S}) = \inf\limits_{P \in \mathcal{S}} W_2^2(X,P)$.*

Proof. For any element $\pi \in \Pi(\mathcal{Q}, \pi_{\mathcal{S}}\#\mathcal{Q})$:

$$\begin{aligned} \int W_2^2(P,G)d\pi(P,G) &\geq \int d^2_{W_2}(P,\mathcal{S})d\pi(P,G) \\ &= \int d^2_{W_2}(P,\mathcal{S})d\mathcal{Q}(P) \\ &= E_{X\sim\mathcal{Q}}(d^2_{W_2}(X,\mathcal{S})) \end{aligned}$$

where the integrations in the first two terms range over $\mathcal{P}_2(\Theta) \times \mathcal{S}$ while that in the final term ranges over $\mathcal{P}_2(\Theta)$. Therefore, we obtain

$$\begin{aligned} W_2^2(\mathcal{Q},\pi_{\mathcal{S}}\#\mathcal{Q}) &= \inf \int\limits_{\mathcal{P}_2(\Theta)\times\mathcal{S}} W_2^2(P,G)d\pi(P,G) \\ &\geq E_{X\sim\mathcal{Q}}(d^2_{W_2}(X,\mathcal{S})) \qquad (2) \end{aligned}$$

where the infimum in the first equality ranges over all $\pi \in \Pi(\mathcal{Q}, \pi_{\mathcal{S}}\#\mathcal{Q})$.

On the other hand, let $g : \mathcal{P}_2(\Theta) \to \mathcal{P}_2(\Theta) \times \mathcal{S}$ such that $g(P) = (P, \pi_{\mathcal{S}}(P))$ for all $P \in \mathcal{P}_2(\Theta)$. Additionally, let $\mu_{\pi_{\mathcal{S}}} = g\#\mathcal{Q}$, the push-forward measure of $\mathcal{Q}$ under mapping g. It is clear that $\mu_{\pi_{\mathcal{S}}}$ is a coupling between $\mathcal{Q}$ and $\pi_{\mathcal{S}}\#\mathcal{Q}$. Under this construction, we obtain for any $X \sim \mathcal{Q}$ that

$$\begin{aligned} E\left(W_2^2(X,\pi_{\mathcal{S}}(X))\right) &= \int W_2^2(P,G)d\mu_{\pi_{\mathcal{S}}}(P,G) \\ &\geq \inf \int W_2^2(P,G)d\pi(P,G) \\ &= W_2^2(\mathcal{Q},\pi_{\mathcal{S}}\#\mathcal{Q}) \qquad (3) \end{aligned}$$

where the infimum in the second inequality ranges over all $\pi \in \Pi(\mathcal{Q}, \pi_{\mathcal{S}}\#\mathcal{Q})$ and the integrations range over $\mathcal{P}_2(\Theta) \times \mathcal{S}$. Now, from the definition of $\pi_{\mathcal{S}}$

$$\begin{aligned} E(W_2^2(X,\pi_{\mathcal{S}}(X))) &= \int W_2^2(P,\pi_{\mathcal{S}}(P))d\mathcal{Q}(P) \\ &= \int d^2_{W_2}(P,\mathcal{S})d\mathcal{Q}(P) \\ &= E(d^2_{W_2}(X,\mathcal{S})) \qquad (4) \end{aligned}$$

where the integrations in the above equations range over $\mathcal{P}_2(\Theta)$. By combining (3) and (4), we would obtain that

$$E_{X\sim\mathcal{Q}}(d^2_{W_2}(X,\mathcal{S})) \geq W_2^2(\mathcal{Q},\pi_{\mathcal{S}}\#\mathcal{Q}). \qquad (5)$$

From (2) and (5), it is straightforward that $E_{X\sim Q}(d(X,S)^2) = W_2^2(Q,\pi_S\#Q)$. Therefore, we achieve the conclusion of the lemma. □

Lemma B.2. *For any closed subset $\mathcal{S} \subset \mathcal{P}_2(\Theta)$ and $\mu \in \mathcal{P}_2(\mathcal{P}_2(\Theta))$ with supp(μ) $\subseteq \mathcal{S}$, there holds $W_2^2(\mathcal{Q},\mu) \geq W_2^2(\mathcal{Q},\pi_{\mathcal{S}}\#\mathcal{Q})$ for any $\mathcal{Q} \in \mathcal{P}_2(\mathcal{P}_2(\Theta))$.*

Proof. Since supp$(\mu) \subseteq \mathcal{S}$, it is clear that $W_2^2(\mathcal{Q},\mu) = \inf\limits_{\pi\in\Pi(\mathcal{Q},\mu)} \int\limits_{\mathcal{P}_2(\Theta)\times\mathcal{S}} W_2^2(P,G)d\pi(P,G)$.

Additionally, we have

$$\begin{aligned} \int W_2^2(P,G)d\pi(P,G) &\geq \int d^2_{W_2}(P,\mathcal{S})d\pi(P,G) \\ &= \int d^2_{W_2}(P,\mathcal{S})d\mathcal{Q}(P) \\ &= E_{X\sim Q}(d^2_{W_2}(X,S)) \\ &= W_2^2(\mathcal{Q},\pi_{\mathcal{S}}\#\mathcal{Q}) \end{aligned}$$

where the last inequality is due to Lemma B.1 and the integrations in the first two terms range over $\mathcal{P}_2(\Theta) \times \mathcal{S}$ while that in the final term ranges over $\mathcal{P}_2(\Theta)$. Therefore, we achieve the conclusion of the lemma. □

Equipped with Lemma B.1 and Lemma B.2, we are ready to establish the equivalence between multilevel Wasserstein means objective function (5) and objective function (4) in Section 3.1 in the main text.

Lemma B.3. *For any given positive integers m and M, we have*

$$A := \inf_{\mathcal{H}\in\mathcal{E}_M(\mathcal{P}_2(\Theta))} W_2^2(\mathcal{H}, \frac{1}{m}\sum_{j=1}^{m}\delta_{G_j})$$

$$= \frac{1}{m}\inf_{\boldsymbol{H}=(H_1,\ldots,H_M)} \sum_{j=1}^{m} d^2_{W_2}(G_j,\boldsymbol{H}) := B.$$

Proof. Write $\mathcal{Q} = \frac{1}{m}\sum\limits_{j=1}^{m} \delta_{G_j}$. From the definition of B, for any $\epsilon > 0$, we can find $\overline{\boldsymbol{H}}$ such that

$$\begin{aligned} B &\geq \frac{1}{m}\sum_{j=1}^{m} d^2_{W_2}(G_j,\overline{\boldsymbol{H}}) - \epsilon \\ &= E_{X\sim\mathcal{Q}}(d^2_{W_2}(X,\overline{\boldsymbol{H}})) - \epsilon \\ &= W_2^2(\mathcal{Q},\pi_{\overline{\boldsymbol{H}}}\#\mathcal{Q}) - \epsilon \\ &\geq A - \epsilon \end{aligned}$$

where the second equality in the above display is due to Lemma B.1 while the last inequality is from the fact that $\pi_{\overline{\boldsymbol{H}}}\#\mathcal{Q}$ is a discrete probability measure in $\mathcal{P}_2(\mathcal{P}_2(\Theta))$ with exactly M support points. Since the inequality in the above display holds for any ϵ, it implies that $B \geq A$. On the other hand, from the formation of A, for any $\epsilon > 0$, we also can find $\mathcal{H}' \in \mathcal{E}_M(\mathcal{P}_2(\Theta))$ such that

$$\begin{aligned} A &\geq W_2^2(\mathcal{H}',\mathcal{Q}) - \epsilon \\ &\geq W_2^2(\mathcal{Q},\pi_{\boldsymbol{H}'}\#\mathcal{Q}) - \epsilon \\ &= \frac{1}{m}\sum_{j=1}^{m} d^2_{W_2}(G_j,\boldsymbol{H}') - \epsilon \\ &\geq B - \epsilon \end{aligned}$$

where $\boldsymbol{H}' = \text{supp}(\mathcal{H}')$, the second inequality is due to Lemma B.2, and the third equality is due to Lemma B.1. Therefore, it means that $A \geq B$. We achieve the conclusion of the lemma. □

Proposition B.4. *For any positive integer numbers m, M and k_j as $1 \leq j \leq m$, we denote*

$$\begin{aligned} C &:= \inf_{\substack{G_j\in\mathcal{O}_{k_j}(\Theta)\ \forall 1\leq j\leq m,\\ \mathcal{H}\in\mathcal{E}_M(\mathcal{P}_2(\Theta))}} \sum_{i=1}^{m} W_2^2(G_j,P^j_{n_j}) \\ &+ W_2^2(\mathcal{H},\frac{1}{m}\sum_{i=1}^{m}\delta_{G_i}) \\ D &:= \inf_{\substack{G_j\in\mathcal{O}_{k_j}(\Theta)\ \forall 1\leq j\leq m,\\ \boldsymbol{H}=(H_1,\ldots,H_M)}} \sum_{j=1}^{m} W_2^2(G_j,P^j_{n_j}) \\ &+ \frac{d^2_{W_2}(G_j,\boldsymbol{H})}{m}. \end{aligned}$$

Then, we have $C = D$.

Proof. The proof of this proposition is a straightforward application of Lemma B.3. Indeed, for each fixed $(G_1,\ldots,G_m)$ the infimum w.r.t to $\mathcal{H}$ in C leads to the same infimum w.r.t to $\boldsymbol{H}$ in D, according to Lemma B.3. Now, by taking the infimum w.r.t to $(G_1,\ldots,G_m)$ on both sides, we achieve the conclusion of the proposition. □

In the remainder of the Supplement, we present the proofs for all remaining theorems stated in the main text.

PROOF OF THEOREM 3.1 The proof of this theorem is straightforward from the formulation of Algorithm 1. In fact, for any $G_j \in \mathcal{E}_{k_j}(\Theta)$ and $\boldsymbol{H} = (H_1,\ldots,H_M)$, we denote the function

$$f(\boldsymbol{G},\boldsymbol{H}) = \sum_{j=1}^{m} W_2^2(G_j,P^j_n) + \frac{d^2_{W_2}(G_j,\boldsymbol{H})}{m}$$

where $\boldsymbol{G} = (G_1,\ldots,G_m)$. To obtain the conclusion of this theorem, it is sufficient to demonstrate for any $t \geq 0$ that

$$f(\boldsymbol{G}^{(t+1)},\boldsymbol{H}^{(t+1)}) \leq f(\boldsymbol{G}^{(t)},\boldsymbol{H}^{(t)}).$$

This inequality comes directly from $f(\boldsymbol{G}^{(t+1)},\boldsymbol{H}^{(t)}) \leq f(\boldsymbol{G}^{(t)},\boldsymbol{H}^{(t)})$, which is due to the Wasserstein barycenter problems to obtain $G_j^{(t+1)}$ for $1 \leq j \leq m$, and $f(\boldsymbol{G}^{(t+1)},\boldsymbol{H}^{(t+1)}) \leq f(\boldsymbol{G}^{(t+1)},\boldsymbol{H}^{(t)})$, which is due to the optimization steps to achieve elements $H_u^{(t+1)}$ of $\boldsymbol{H}^{(t+1)}$ as $1 \leq u \leq M$. As a consequence, we achieve the conclusion of the theorem.

PROOF OF THEOREM 4.1 To simplify notation, write

$$\begin{aligned} L_{\boldsymbol{n}} &= \inf_{\substack{G_j\in\mathcal{O}_{k_j}(\Theta),\\ \mathcal{H}\in\mathcal{E}_M(\mathcal{P}_2(\Theta))}} f_{\boldsymbol{n}}(\boldsymbol{G},\mathcal{H}), \\ L_0 &= \inf_{\substack{G_j\in\mathcal{O}_{k_j}(\Theta),\\ \mathcal{H}\in\mathcal{E}_M(\mathcal{P}_2(\Theta))}} f(\boldsymbol{G},\mathcal{H}). \end{aligned}$$

For any $\epsilon > 0$, from the definition of L_0, we can find $G_j \in \mathcal{O}_{k_j}(\Theta)$ and $\mathcal{H} \in \mathcal{E}_M(\mathcal{P}(\Theta))$ such that

$$f(\boldsymbol{G},\mathcal{H})^{1/2} \leq L_0^{1/2} + \epsilon.$$

Therefore, we would have

$$\begin{aligned} L_{\boldsymbol{n}}^{1/2} - L_0^{1/2} &\leq L_n^{1/2} - f(\boldsymbol{G},\mathcal{H})^{1/2} + \epsilon \\ &\leq f_{\boldsymbol{n}}(\boldsymbol{G},\mathcal{H})^{1/2} - f(\boldsymbol{G},\mathcal{H})^{1/2} + \epsilon \\ &= \frac{f_{\boldsymbol{n}}(\boldsymbol{G},\mathcal{H}) - f(\boldsymbol{G},\mathcal{H})}{f_{\boldsymbol{n}}(\boldsymbol{G},\mathcal{H})^{1/2} + f(\boldsymbol{G},\mathcal{H})^{1/2}} + \epsilon \\ &\leq \sum_{j=1}^{m} \frac{|W_2^2(G_j,P^j_{n_j}) - W_2^2(G_j,P^j)|}{W_2(G_j,P^j_{n_j}) + W_2(G_j,P^j)} + \epsilon \\ &\leq \sum_{j=1}^{m} W_2(P^j_{n_j},P^j) + \epsilon. \end{aligned}$$

By reversing the direction, we also obtain the inequality $L_n^{1/2} - L_0^{1/2} \geq \sum\limits_{j=1}^{m} W_2(P^j_{n_j},P^j) - \epsilon$. Hence, $|L_n^{1/2} - L_0^{1/2} - \sum\limits_{j=1}^{m} W_2(P^j_{n_j},P^j)| \leq \epsilon$ for any $\epsilon > 0$. Since $P^j \in \mathcal{P}_2(\Theta)$ for all $1 \leq j \leq m$, we obtain that $W_2(P^j_{n_j},P^j) \to 0$ almost surely as $n_j \to \infty$ (see for example Theorem 6.9 in (Villani, 2009)). As a consequence, we obtain the conclusion of the theorem.

PROOF OF THEOREM 4.2 For any $\epsilon > 0$, we denote

$$\mathcal{A}(\epsilon) = \Big\{ G_i \in \mathcal{O}_{k_i}(\Theta), \mathcal{H} \in \mathcal{E}_M(\mathcal{P}(\Theta)) : \\ d(\boldsymbol{G}, \mathcal{H}, \mathcal{F}) \geq \epsilon \Big\}.$$

Since Θ is a compact set, we also have $\mathcal{O}_{k_j}(\Theta)$ and $\mathcal{E}_M(\mathcal{P}_2(\Theta))$ are compact for any $1 \leq i \leq m$. As a consequence, $\mathcal{A}(\epsilon)$ is also a compact set. For any $(\boldsymbol{G}, \mathcal{H}) \in \mathcal{A}(\epsilon)$, by the definition of $\mathcal{F}$ we would have $f(\boldsymbol{G}, \mathcal{H}) > f(\boldsymbol{G}^0, \mathcal{H}^0)$ for any $(\boldsymbol{G}^0, \mathcal{H}^0) \in \mathcal{F}$. Since $\mathcal{A}(\epsilon)$ is compact, it leads to

$$\inf_{(\boldsymbol{G}, \mathcal{H}) \in A(\epsilon)} f(\boldsymbol{G}, \mathcal{H}) > f(\boldsymbol{G}^0, \mathcal{H}^0).$$

for any $(\boldsymbol{G}^0, \mathcal{H}^0) \in \mathcal{F}$. From the formulation of $f_{\boldsymbol{n}}$ as in the proof of Theorem 4.1, we can verify that $\lim\limits_{\boldsymbol{n} \to \infty} f_{\boldsymbol{n}}(\widehat{\boldsymbol{G}}^{\boldsymbol{n}}, \widehat{\mathcal{H}}^{\boldsymbol{n}}) = \lim\limits_{\boldsymbol{n} \to \infty} f(\widehat{\boldsymbol{G}}^{\boldsymbol{n}}, \widehat{\mathcal{H}}^{\boldsymbol{n}})$ almost surely as $\boldsymbol{n} \to \infty$. Combining this result with that of Theorem 4.1, we obtain $f(\widehat{\boldsymbol{G}}^{\boldsymbol{n}}, \widehat{\mathcal{H}}^{\boldsymbol{n}}) \to f(\boldsymbol{G}^0, \mathcal{H}^0)$ as $\boldsymbol{n} \to \infty$ for any $(\boldsymbol{G}^0, \mathcal{H}^0) \in \mathcal{F}$. Therefore, for any $\epsilon > 0$, as $\boldsymbol{n}$ is large enough, we have $d(\widehat{\boldsymbol{G}}^{\boldsymbol{n}}, \widehat{\mathcal{H}}^{\boldsymbol{n}}, \mathcal{F}) < \epsilon$. As a consequence, we achieve the conclusion regarding the consistency of the mixing measures.

Appendix C

In this appendix, we provide details on the algorithm for the Multilevel Wasserstein means with sharing (MWMS) formulation (Algorithm 2). Recall the MWMS objective function as follows

$$\inf_{\mathcal{S}_K, G_j, \boldsymbol{H} \in \mathcal{B}_{M, \mathcal{S}_K}} \sum_{j=1}^{m} W_2^2(G_j, P_{n_j}^j) + \frac{d_{W_2}^2(G_j, \boldsymbol{H})}{m}$$

where $\mathcal{B}_{M, \mathcal{S}_K} = \Big\{ G_j \in \mathcal{O}_K(\Theta),\ \boldsymbol{H} = (H_1, \ldots, H_M) :$ supp$(G_j) \subseteq \mathcal{S}_K\ \forall 1 \leq j \leq m \Big\}$.

We make the following remarks regarding the initializations and updates of Algorithm 2:

(i) An efficient way to initialize global set $S_K^{(0)} = \Big\{ a_1^{(0)}, \ldots, a_K^{(0)} \Big\} \in \mathbb{R}^{d \times K}$ is to perform K-means on the whole data set $X_{j,i}$ for $1 \leq j \leq m, 1 \leq i \leq n_j$;

(ii) The updates $a_j^{(t+1)}$ are indeed the solutions of the following optimization problems

$$\inf_{a_j^{(t)}} \Bigg\{ \sum_{l=1}^{m} W_2^2(G_l^{(t)}, P_n^l) + \frac{\sum\limits_{l=1}^{m} W_2^2(G_l^{(t)}, H_{i_l}^{(t)})}{m} \Bigg\},$$

Algorithm 2 Multilevel Wasserstein Means with Sharing (MWMS)

```
Input: Data $X_{j,i}$, $K$, $M$.
Output: global set $S_K$, local measures $G_j$, and elements $H_i$ of $\boldsymbol{H}$.
Initialize $S_K^{(0)} = \Big\{ a_1^{(0)}, \ldots, a_K^{(0)} \Big\}$, elements $H_i^{(0)}$ of $\boldsymbol{H}^{(0)}$, and $t = 0$.
while $S_K^{(t)}, G_j^{(t)}, H_i^{(t)}$ have not converged do
    1. Update global set $S_K^{(t)}$:
    for $j = 1$ to $m$ do
        $i_j \leftarrow \arg\min\limits_{1 \leq u \leq M} W_2^2(G_j^{(t)}, H_u^{(t)})$.
        $T^j \leftarrow$ optimal coupling of $G_j^{(t)}$, $P_n^j$ (cf. Appendix A).
        $U^j \leftarrow$ optimal coupling of $G_j^{(t)}$, $H_{i_j}^{(t)}$.
    end for
    for $i = 1$ to $M$ do
        $h_i^{(t)} \leftarrow$ atoms of $H_i^{(t)}$ with $h_{i,v}^{(t)}$ as v-th column.
    end for
    for $i = 1$ to $K$ do
        $mD \leftarrow m \sum\limits_{u=1}^{m} \sum\limits_{v=1}^{n_i} T_{i,v}^u + \sum\limits_{u=1}^{m} \sum\limits_{v \neq i} U_{i,v}^u$.
        $a_i^{(t+1)} \leftarrow \Big( m \sum\limits_{u=1}^{m} \sum\limits_{v=1}^{n_i} T_{i,v}^u X_{u,v} +$
        $\sum\limits_{u=1}^{m} \sum\limits_{v} U_{i,v}^u h_{j_u,v}^{(t)} \Big) / mD$.
    end for
    2. Update $G_j^{(t)}$ for $1 \leq j \leq m$:
    for $j = 1$ to $m$ do
        $G_j^{(t+1)} \leftarrow \arg\min\limits_{G_j : \text{supp}(G_j) \equiv \mathcal{S}_K^{(t+1)}} W_2^2(G_j, P_{n_j}^j)$
        $+ W_2^2(G_j, H_{i_j}^{(t)}) / m$.
    end for
    3. Update $H_i^{(t)}$ for $1 \leq i \leq M$ as Algorithm 1.
    4. $t \leftarrow t + 1$.
end while
```

which is equivalent to find $a_j^{(t)}$ to optimize

$$m\sum_{u=1}^{m}\sum_{v=1}^{n_j} T_{j,v}^{u}\|a_j^{(t)} - X_{u,v}\|^2 + \sum_{u=1}^{m}\sum_{v} U_{j,v}^{u}\|a_j^{(t)} - h_{i_j,v}^{(t)}\|^2.$$

where T^j is an optimal coupling of $G_j^{(t)}$, P_n^j and U^j is an optimal coupling of $G_j^{(t)}$, $H_{i_j}^{(t)}$. By taking the first order derivative of the above function with respect to $a_j^{(t)}$, we quickly achieve $a_j^{(t+1)}$ as the closed form minimum of that function;

(iii) Updating the local weights of $G_j^{(t+1)}$ is equivalent to updating $G_j^{(t+1)}$ as the atoms of $G_j^{(t+1)}$ are known to stem from $S_K^{(t+1)}$.

Now, similar to Theorem 3.1 in the main text, we also have the following theoretical guarantee regarding the behavior of Algorithm 2 as follows

Theorem C.1. *Algorithm 2 monotonically decreases the objective function of the MWMS formulation.*

Proof. The proof is quite similar to the proof of Theorem 3.1. In fact, recall from the proof of Theorem 3.1 that for any $G_j \in \mathcal{E}_{k_j}(\Theta)$ and $\boldsymbol{H} = (H_1, \ldots, H_M)$ we denote the function

$$f(\boldsymbol{G}, \boldsymbol{H}) = \sum_{j=1}^{m} W_2^2(G_j, P_n^j) + \frac{d_{W_2}^2(G_j, \boldsymbol{H})}{m}$$

where $\boldsymbol{G} = (G_1, \ldots, G_m)$. Now it is sufficient to demonstrate for any $t \geq 0$ that

$$f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)}) \leq f(\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)}).$$

where the formulation of f is similar as in the proof of Theorem 3.1. Indeed, by the definition of Wasserstein distances, we have

$$E = mf(\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)}) = \sum_{u=1}^{m}\sum_{j,v} mT_{j,v}^{u}\|a_j^{(t)} - X_{u,v}\|^2 + U_{j,v}^{u}\|a_j^{(t)} - h_{i_u,v}^{(t)}\|^2.$$

Therefore, the update of $a_i^{(t+1)}$ from Algorithm 2 leads to

$$\begin{aligned} E &\geq \sum_{u=1}^{m}\sum_{j,v} mT_{j,v}^{u}\|a_j^{(t+1)} - X_{u,v}\|^2 \\ &+ U_{j,v}^{u}\|a_j^{(t+1)} - h_{i_u,v}^{(t)}\|^2 \\ &\geq m\sum_{j=1}^{m} W_2^2(G_j^{(t)'}, P_n^j) + \sum_{j=1}^{m} W_2^2(G_j^{(t)'}, H_{i_j}^{(t)}) \\ &\geq m\sum_{j=1}^{m} W_2^2(G_j^{(t)'}, P_n^j) + \sum_{j=1}^{m} d_{W_2}^2(G_j^{(t)'}, \boldsymbol{H}^{(t)}) \\ &= mf(\boldsymbol{G'}^{(t)}, \boldsymbol{H}^{(t)}) \end{aligned}$$

where $\boldsymbol{G'}^{(t)} = (G_1^{(t)'}, \ldots, G_m^{(t)'})$, $G_j^{(t)'}$ are formed by replacing the atoms of $G_j^{(t)}$ by the elements of $S_K^{(t+1)}$, noting that supp$(G_j^{(t)'}) \subseteq \mathcal{S}_K^{(t+1)}$ as $1 \leq j \leq m$, and the second inequality comes directly from the definition of Wasserstein distance. Hence, we obtain

$$f(\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)}) \geq f(\boldsymbol{G'}^{(t)}, \boldsymbol{H}^{(t)}). \tag{6}$$

From the formation of $G_j^{(t+1)}$ as $1 \leq j \leq m$, we get

$$\sum_{j=1}^{m} d_{W_2}^2(G_j^{(t+1)}, \boldsymbol{H}^{(t)}) \leq \sum_{j=1}^{m} d_{W_2}^2(G_j^{(t)'}, \boldsymbol{H}^{(t)}).$$

Thus, it leads to

$$f(\boldsymbol{G'}^{(t)}, \boldsymbol{H}^{(t)}) \geq f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t)}). \tag{7}$$

Finally, from the definition of $H_1^{(t+1)}, \ldots, H_M^{(t+1)}$, we have

$$f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t)}) \geq f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)}). \tag{8}$$

By combining (6), (7), and (8), we arrive at the conclusion of the theorem. □

Appendix D

In this appendix, we offer details on the data generation processes utilized in the simulation studies presented in Section 5 in the main text. The notions of m, n, d, M are given in the main text. Let K_i be the number of supporting atoms of H_i and k_j the number of atoms of G_j. For any $d \geq 1$, we denote $\mathbf{1}_d$ to be d dimensional vector with all components to be 1. Furthermore, $\mathcal{I}_d$ is an identity matrix with d dimensions.

Comparison metric (Wasserstein distance to truth)

$$\mathrm{W} := \frac{1}{m}\sum_{j=1}^{m} W_2(\hat{G}_j, G_j) + d_{\mathcal{M}}(\hat{\boldsymbol{H}}, \boldsymbol{H})$$

where $\hat{\boldsymbol{H}} := \{\hat{H}_1, \ldots, \hat{H}_M\}$, $\boldsymbol{H} := \{H_1, \ldots, H_M\}$ and $d_{\mathcal{M}}(\hat{H}, H)$ is a minimum-matching distance (Tang et al., 2014; Nguyen, 2015):

$$d_{\mathcal{M}}(\hat{\boldsymbol{H}}, \boldsymbol{H}) := \max\{\overline{d}(\hat{\boldsymbol{H}}, \boldsymbol{H}), \overline{d}(\boldsymbol{H}, \hat{\boldsymbol{H}})\}$$

where

$$\overline{d}(\hat{\boldsymbol{H}}, \boldsymbol{H}) := \max_{1 \le i \le M} \min_{1 \le j \le M} W_2(H_i, \hat{H}_j).$$

Multilevel Wasserstein means setting The global clusters are generated as follows:

means for atoms $\mu_i := 5(i-1), i = 1, \ldots, M$.
atoms of $H_i : \phi_{ij} \sim \mathcal{N}(\mu_i \mathbf{1}_d, \mathcal{I}_d), j = 1, \ldots, K_i$.
weights of atoms: $\pi_i \sim \text{Dir}(\mathbf{1}_{K_i})$.
Let $H_i := \sum_{j=1}^{K_i} \pi_{ij} \delta_{\phi_{ij}}$.

For each group $j = 1, \ldots, m$, generate local measures and data as follows:

pick cluster label $z_j \sim \text{Unif}(\{1, \ldots, M\})$.
mean for atoms : $\tau_{ji} \sim H_{z_j}, i = 1, \ldots, k_j$.
atoms of $G_j : \theta_{ji} \sim \mathcal{N}(\tau_{ji}, \mathcal{I}_d), i = 1, \ldots, k_j$.
weights of atoms $p_j \sim \text{Dir}(\mathbf{1}_{k_j})$.
Let $G_j := \sum_{i=1}^{k_j} p_{ji} \delta_{\theta_{ji}}$.
data mean $\mu_i \sim G_j, i = 1, \ldots, n_j$.
observation $X_{j,i} \sim \mathcal{N}(\mu_i, \mathcal{I}_d)$.

For the case of non-constrained variances, the variance to generate atoms θ_{ji} of G_j is set to be proportional to global cluster label z_j assigned to G_j.

Multilevel Wasserstein means with sharing setting The global clusters are generated as follows:

means for atoms $\mu_i := 5(i-1), i = 1, \ldots, M$.
atoms of $H_i : \phi_{ij} \sim \mathcal{N}(\mu_i \mathbf{1}_d, \mathcal{I}_d), j = 1, \ldots, K_i$.
weights of atoms $\pi_i \sim \text{Dir}(\mathbf{1}_{K_i})$.
Let $H_i := \sum_{j=1}^{K_i} \pi_{ij} \delta_{\phi_{ij}}$.

For each shared atom $k = 1, \ldots, K$:

pick cluster label $z_k \sim \text{Unif}(\{1, \ldots, M\})$.
mean for atoms : $\tau_k \sim H_{z_k}$.
atoms of $S_K : \theta_k \sim \mathcal{N}(\tau_k, \mathcal{I}_d)$.

For each group $j = 1, \ldots, m$ generate local measures and data as follows:

pick cluster label $\tilde{z}_j \sim \text{Unif}(\{1, \ldots, M\})$.
select shared atoms $s_j = \{k : z_k = \tilde{z}_j\}$.
weights of atoms $p_{s_j} \sim \text{Dir}(\mathbf{1}_{|s_j|})$; $\quad G_j := \sum_{i \in s_j} p_i \delta_{\theta_i}$.
data mean $\mu_i \sim G_j, i = 1, \ldots, n_j$.
observation $X_{j,i} \sim \mathcal{N}(\mu_i, \mathcal{I}_d)$.

For the case of non-constrained variances, the variance to generate atoms θ_i of G_j where $i \in s_j$ is set to be proportional to global cluster label $\tilde{z}_j$ assigned to G_j.

Three-stage K-means First, we estimate G_j for each group $1 \le j \le m$ by using K-means algorithm with k_j clusters. Then, we cluster labels using K-means algorithm with M clusters based on the collection of all atoms of G_js. Finally, we estimate the atoms of each H_i via K-means algorithm with exactly L clusters for each group of local atoms. Here, L is some given threshold being used in Algorithm 1 in Section 3.1 in the main text to speed up the computation (see final remark regarding Algorithm 1 in Section 3.1). The three-stage K-means algorithm is summarized in Algorithm 3.

Algorithm 3 Three-stage K-means

```
Input: Data $X_{j,i}$, $k_j$, $M$, $L$.
Output: local measures $G_j$ and global elements $H_i$ of $\boldsymbol{H}$.
Stage 1
for $j = 1$ to $m$ do
    $G_j \leftarrow k_j$ clusters of group j with K-means (atoms as centroids and weights as label frequencies).
end for
$\mathcal{C} \leftarrow$ collection of all atoms of $G_j$.
Stage 2
$\{D_1, \ldots, D_M\} \leftarrow M$ clusters from K-means on $\mathcal{C}$.
Stage 3
for $i = 1$ to $M$ do
    $H_i \leftarrow L$ clusters of $D_i$ with K-means (atoms as centroids and weights as label frequencies).
end for
```

References

M. Agueh and G. Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43: 904–924, 2011.

E. Anderes, S. Borgwardt, and J. Miller. Discrete wasserstein barycenters: optimal transport for discrete data. *http://arxiv.org/abs/1507.07218*, 2015.

M. Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems 26*, 2013.

X. Nguyen. Posterior contraction of the population polytope in finite admixture models. *Bernoulli*, 21:618–646, 2015.

Jian Tang, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of The 31st International Conference on Machine Learning*, pages 190–198. ACM, 2014.

C. Villani. *Optimal Transport: Old and New. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathemtical Sciences]*. Springer, Berlin, 2009.