Projection Robust Wasserstein Barycenters

Minhui Huang [1] Shiqian Ma [2] Lifeng Lai [1]

[1]Department of Electrical and Computer Engineering
[2]Department of Mathematics
University of California, Davis

June 16, 2021

Wasserstein Barycenter

- Aggregating information from several probability measures or histograms.

Wasserstein Barycenter

- Aggregating information from several probability measures or histograms.
- 2-Wasserstein distance between measures $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$:

$$\mathcal{W}^2(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^2 d\pi(x, y). \tag{1}$$

Wasserstein Barycenter

- Aggregating information from several probability measures or histograms.
- 2-Wasserstein distance between measures $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$:

$$\mathcal{W}^2(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^2 d\pi(x, y). \tag{1}$$

- The Wasserstein Barycenter of m probability measures $\boldsymbol{\mu} := \{\mu^l\}_{l \in [m]}$:

$$\inf_{\nu \in \mathscr{P}_2(\mathbb{R}^d)} \mathcal{WB}(\boldsymbol{\mu}, \boldsymbol{\omega}) := \sum_{l=1}^{m} \omega^l \mathcal{W}^2(\mu^l, \nu). \tag{2}$$

Projection Robust Wasserstein Barycenter

- (Paty and Cuturi, 2019) Projection Robust Wasserstein (PRW) distance:

$$\mathcal{P}_k(\mu, \nu) := \sup_{E \in \mathcal{G}_k} \mathcal{W}(\mathrm{Proj}_E \mu, \mathrm{Proj}_E \nu). \tag{3}$$

Projection Robust Wasserstein Barycenter

- (Paty and Cuturi, 2019) Projection Robust Wasserstein (PRW) distance:

$$\mathcal{P}_k(\mu, \nu) := \sup_{E \in \mathcal{G}_k} \mathcal{W}(\mathrm{Proj}_E \mu, \mathrm{Proj}_E \nu). \tag{3}$$

- Projection Robust Wasserstein Barycenter:

$$\begin{aligned} &\inf_{\nu \in \mathscr{P}_2(\mathbb{R}^d)} \sum_{l=1}^{m} \omega^l \mathcal{P}_k^2(\mu^l, \nu) \\ = &\inf_{\nu \in \mathscr{P}_2(\mathbb{R}^d)} \sum_{l=1}^{m} \omega^l \sup_{U_\ell \in \mathrm{St}(d,k)} \inf_{\pi^l \in \Pi(\mu^l, \nu)} \int \| U_\ell^\top (x^l - y) \|^2 d\pi^l(x^l, y). \end{aligned} \tag{4}$$

An inf-sup-inf problem.

A Relaxed Model

- Two relaxations:
  - Use a common projector $\mathrm{Proj}_E(\cdot)$ for all PRW distances.
  - Switch the order of sup and the first inf.

A Relaxed Model

- Two relaxations:
  - Use a common projector $\mathrm{Proj}_E(\cdot)$ for all PRW distances.
  - Switch the order of sup and the first inf.
- Relaxed PRWB

$$\begin{aligned} &\sup_{E\in\mathcal{G}_k}\inf_{\nu\in\mathscr{P}_2(\mathbb{R}^d)}\sum_{l=1}^{m}\omega^l\mathcal{W}^2(\mathrm{Proj}_E\mu^l,\mathrm{Proj}_E\nu) \\ =&\sup_{U\in\mathrm{St}(d,k)}\inf_{\nu\in\mathscr{P}_2(\mathbb{R}^d)}\sum_{l=1}^{m}\omega^l\inf_{\pi^l\in\Pi(\mu^l,\nu)}\int\|U^\top(x^l-y)\|^2d\pi^l(x^l,y). \end{aligned} \tag{5}$$

A Relaxed Model

- Two relaxations:
  - Use a common projector $\mathrm{Proj}_E(\cdot)$ for all PRW distances.
  - Switch the order of sup and the first inf.
- Relaxed PRWB

$$
\begin{aligned}
&\sup_{E \in \mathcal{G}_k} \inf_{\nu \in \mathscr{P}_2(\mathbb{R}^d)} \sum_{l=1}^{m} \omega^l \mathcal{W}^2(\mathrm{Proj}_E \mu^l, \mathrm{Proj}_E \nu) \\
= &\sup_{U \in \mathrm{St}(d,k)} \inf_{\nu \in \mathscr{P}_2(\mathbb{R}^d)} \sum_{l=1}^{m} \omega^l \inf_{\pi^l \in \Pi(\mu^l, \nu)} \int \|U^\top (x^l - y)\|^2 d\pi^l(x^l, y).
\end{aligned} \tag{5}
$$

- Entropy regularized model:

$$
\max_{U \in \mathcal{M}} \min_{\pi \in \Pi(\boldsymbol{p})} f_\eta(\boldsymbol{\pi}, U) := \sum_{l=1}^{m} \omega^l \sum_{i,j=1}^{n} \pi^l_{i,j} \|U^\top (x^l_i - y_j)\|^2 - \eta H(\pi^l). \tag{6}
$$

Two Algorithms

- RGA-IBP
  - Algorithm update:
    - Solve $f_\eta(U) = \min_{\boldsymbol{\pi}} f_\eta(\boldsymbol{\pi}, U)$;
    - $U_{t+1} := \mathrm{Retr}_{U_t}(\tau \mathrm{grad}\, f_\eta(U_t))$
  - Convergence rate: $O(mn^2 d\epsilon^{-4} + mn^2\epsilon^{-10})$

Two Algorithms

- RGA-IBP
  - Algorithm update:
    - Solve $f_\eta(U) = \min_{\boldsymbol{\pi}} f_\eta(\boldsymbol{\pi}, U)$;
    - $U_{t+1} := \mathrm{Retr}_{U_t}(\tau \mathrm{grad}\, f_\eta(U_t))$
  - Convergence rate: $O(mn^2 d\epsilon^{-4} + mn^2\epsilon^{-10})$
- RBCD
  - Derive the dual: $\min_{\substack{\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^{m\times n},\ U\in\mathcal{M},\\ \sum_{l=1}^m \omega^l v^l = 0}} g(\boldsymbol{u}, \boldsymbol{v}, U) :=$
    $-\sum_{l=1}^m \omega^l \left\{\sum_{i,j=1}^n \exp\left(-\frac{\|U^\top(x_i^l - y_j)\|^2}{\eta} + u_i^l + v_j^l\right) - \langle u^l, p^l\rangle\right\}$
  - Algorithm update:
    - $\boldsymbol{u}_{t+1} = \mathrm{argmin}_{\boldsymbol{u}}\, g(\boldsymbol{u}, \boldsymbol{v}_t, U_t)$;
    - $\boldsymbol{v}_{t+1} = \mathrm{argmin}_{\boldsymbol{v}:\sum_{l=1}^m \omega^l v^l = 0}\, g(\boldsymbol{u}_{t+1}, \boldsymbol{v}, U_t)$;
    - $U_{t+1} = \mathrm{Retr}_{U_t}(-\tau \mathrm{grad}_U g(\boldsymbol{u}_{t+1}, \boldsymbol{v}_{t+1}, U_t))$;
  - Convergence rate: $O\left(mn^2 d\epsilon^{-3}\right)$.

Projected Discrete Distribution Clustering

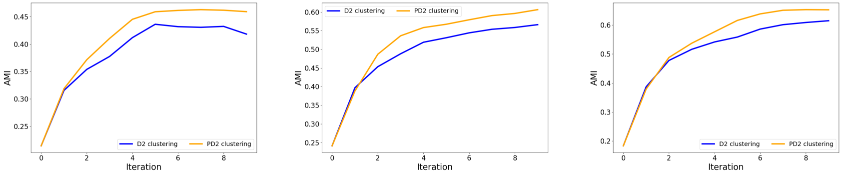

Figure: AMI scores for each iteration. **Left:** the “Reuters Subset” dataset, **Middle:** the “BBCsport Abstract” dataset, **Right:** the “BBCnews Abstract” dataset. The results are averaged on 5 runs.