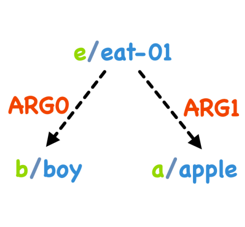
Exploiting Edge Features in Graph-based Learning with Fused Network Gromov-Wasserstein Distance

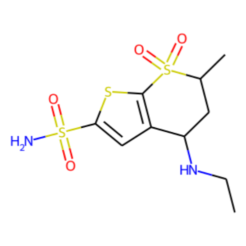
Junjie Yang, Matthieu Labeau, Florence d'Alché-Buc

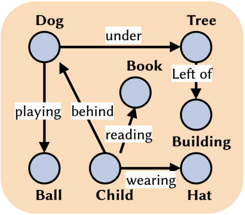
July 18, 2024

LTCI, Télécom Paris, IP Paris

Motivation







Recent advances in OT for graphs has shown to be useful in different graph-based learning tasks:

- Graph Classification [Vayer et al., 2019]
- Graph Clustering [Peyré et al., 2016, Vayer et al., 2019]
- Graph Dictionary Learning [Vincent-Cuaz et al., 2021]
- Supervised Graph Prediction [Brogat-Motte et al., 2022]

Motivation: Unlock OT-based learning for edge featured graphs. We target especially Supervised Graph Prediction problem.

Backgound

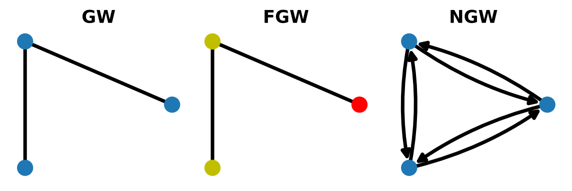


Figure 2: Existing OT-based distance on graph objects: Gromov-Wasserstein [Mémoli, 2011, Sturm, 2012], Fused Gromov-Wasserstein [Vayer et al., 2019], Network Gromov-Wasserstein [Chowdhury and Mémoli, 2019].

Fused Network Gromov-Wasserstein Distance

Node and Edge Featured Graph

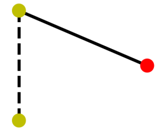
Definition (Node and Edge Featured Graph)

A node and edge featured graph of size m is a quadruple of the form $(F, A, E, \boldsymbol{p})$ where

- $F \in \Psi^m$ is a tuple of points valued in a metric space (Ψ, d_Ψ)
- $A \in \mathbb{R}^{m \times m}$ is a real-valued matrix
- $E \in \Omega^{m \times m}$ is a tuple of points valued in a metric space (Ω, d_Ω)
- $\boldsymbol{p} \in \Sigma_m$ is a simplex histogram

We denote $\mathcal{G}$ as a set of such quadruples.

Node and Edge Featured Graph



Example (Node and Edge Featured Graph)

- $\Psi = \{\text{red}, \text{yellow}\}$: the node-color space
- $\Omega = \{\text{solid}, \text{dashed}, \text{non-edge}\}$: the edge-type space

$$F = \begin{bmatrix} [1,0] \\ [1,0] \\ [0,1] \end{bmatrix}, A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

$$E = \begin{bmatrix} [0,0,1] & [0,1,0] & [1,0,0] \\ [0,1,0] & [0,0,1] & [0,0,1] \\ [1,0,0] & [0,0,1] & [0,0,1] \end{bmatrix}, \boldsymbol{p} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

Fused Network Gromov-Wasserstein Distance

Definition (FNGW Distance, Discrete Case [1])

Given $g = (F, A, E, \boldsymbol{p})$ of size m, $\tilde{g} = (\tilde{F}, \tilde{A}, \tilde{E}, \tilde{\boldsymbol{p}})$ of size $\tilde{m}$ corresponding to two tuples of $\mathcal{G}$, and trade-off parameters $(\alpha, \beta) \in [0, 1]^2$, the Fused Network Gromov-Wasserstein distance between them for $(p, q) \in [1, \infty]$ is written as :

$$\mathrm{FNGW}_{\alpha,\beta,q,p}(g, \tilde{g}) = \min_{\pi \in \Pi(\boldsymbol{p}, \tilde{\boldsymbol{p}})} \mathcal{E}_{\alpha,\beta,q,p}((F, A, E), (\tilde{F}, \tilde{A}, \tilde{E}), \pi) \tag{1}$$

with

$$\mathcal{E}_{\alpha,\beta,q,p}((F, A, E), (\tilde{F}, \tilde{A}, \tilde{E}), \pi) = \Bigg(\sum_{i,j,k,l} \Bigg[\alpha d_\Omega \left(E(i, k), \tilde{E}(j, l) \right)^q$$

$$+\beta |A(i, k) - \tilde{A}(j, l)|^q + (1 - \alpha - \beta) d_\Psi \left(F(i), \tilde{F}(j) \right)^q \Bigg]^p \pi_{k,l} \pi_{i,j} \Bigg)^{\frac{1}{p}} \tag{2}$$

[1] A general definition of FNGW distance is also given in the paper.

Fused Network Gromov-Wasserstein Distance

Example (FNGW Distance)

The FNGW distance between the two graphs illustrated above is 0.296 when $\alpha = \frac{1}{3}$, $\beta = \frac{1}{3}$, $p = 1$ and $q = 2$, where the FGW distance between them is 0.

Computation algorithm: When $p = 1$ and $q = 2$, we adopt Conditional Gradient Descent (CGD) as in [Vayer et al., 2020] to compute FNGW distance.

Metric Properties

The FNGW distance satisfies the following **metric** properties: positivity, symmetry, equality with a corresponding notion of weak isomorphism, relaxed triangle inequality with a factor of 2^{q-1}.

Fused Network Gromov-Wasserstein Barycenter

Fused Network Gromov-Wasserstein Barycenter

Definition (FNGW Barycenter)

Given a set $\{g_i\}_{i=1}^n$ with $g_i = (F_i, A_i, E_i, \boldsymbol{p}_i) \in \mathbb{R}^{m_i \times S} \times \mathbb{R}^{m_i \times m_i} \times \mathbb{R}^{m_i \times m_i \times T} \times \Sigma_{m_i}$ and a set of weights $\{\lambda_i\}_{i=1}^n$ such that $\sum_i \lambda_i = 1$, the FNGW Barycenter for a pre-defined histogram $\boldsymbol{p} \in \Sigma_m$ is defined as follows:

$$\mathfrak{B}(\{\lambda_i\}_i, \{g_i\}_i, \boldsymbol{p}) = \underset{F \in \mathbb{R}^{m \times S}, A \in \mathbb{R}^{m \times m}, E \in \mathbb{R}^{m \times m \times T}}{\arg\min} \sum_i \lambda_i \mathrm{FNGW}_{\alpha,\beta}((F, A, E, \boldsymbol{p}), g_i)$$

Computation algorithm: Block Coordinate Descent.

Proposition

Optimizing above Equation with respect to tensor E has a closed-form solution:

$$E = \frac{1}{\mathcal{I}_{m \times T} \times_2 \boldsymbol{p}\boldsymbol{p}^\top} \sum_i \lambda_i (E_i \times_2 \pi_i) \times_1 \pi_i \tag{3}$$

Property of FNGW Barycenter

Proposition

If the set of tensors $\{E_i\}_i$ satisfies the condition:

$$\forall j, l, i, \sum_{t}^{T} E_i(j, l, t) = a \in \mathbb{R} \tag{4}$$

then the barycenter E given by our alogithm also verify the same property.

One interesting consequence: When the edge labels of the graphs are represented using one-hot encoding, the resulting barycenter can be discretized into a true graph by applying a simple *argmax* operation on the edge features, due to their simplex nature.

$\Rightarrow$ Useful for labeled graph prediction

Examples of FNGW Barycenter

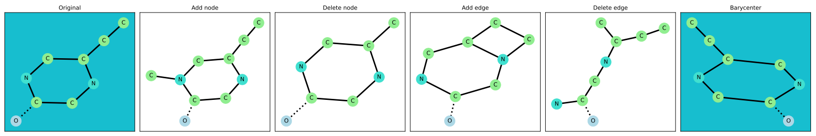


Figure 3: FNGW barycenter (rightmost) of the graphs obtained by perturbing a random molecule (leftmost).

Structured Prediction with FNGW Barycenter

Structured Prediction: Supervised Graph Prediction

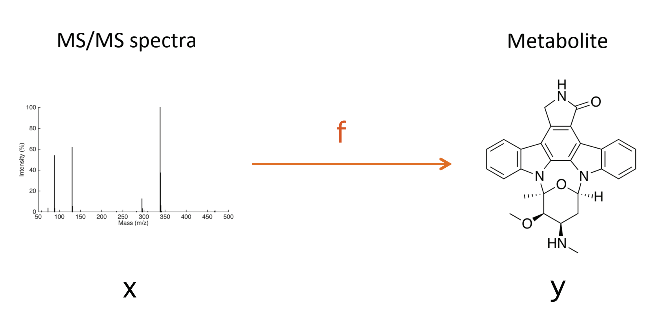


Figure 4: Metabolite Identification Task

Existing works:

- Kernel induced loss [Brouard et al., 2016]
- OT-based loss [Brogat-Motte et al., 2022]

$\Rightarrow$ Surrogate Regression Framework - ILE

(Relaxed) Supervised Graph Prediction

Structured graph space:

$$\mathcal{G} = \big\{(F, A, E, \boldsymbol{p}) \mid m_g \leq m_{\max},\ A \in \{0,1\}^{m_g \times m_g},\ F = (F_i)_{i=1}^{m_g} \in \mathcal{F}^{m_g},$$
$$E = (E_{ij}) \in \mathcal{T}^{m_g \times m_g},\ \boldsymbol{p} = {m_g}^{-1}\mathbb{1}_{m_g}\big\} \quad (5)$$

where $\mathcal{F} \subset \mathbb{R}^S$ and $\mathcal{T} \subset \mathbb{R}^T$ are finite node and edge features spaces.

Relaxed graph space:

$$\mathcal{G}_m = \big\{(F, A, E, \boldsymbol{p}) \mid A \in [0,1]^{m \times m},\ F = (F_i)_{i=1}^{m} \in \mathrm{Conv}(\mathcal{F})^m,$$
$$E = (E_{ij}) \in \mathrm{Conv}(\mathcal{T})^{m \times m},\ \boldsymbol{p} = m^{-1}\mathbb{1}_m\big\} \quad (6)$$

(Relaxed) Supervised Graph Prediction

Given a set of training pairs consisting of inputs and graphs to be predicted $\{(x_i, g_i)\}_{i=1}^n$ drawn from a fixed but unknown distribution ρ on $\mathcal{X} \times \mathcal{G}$.

We are interested in the relaxed supervised graph prediction problem, i.e., finding an estimator $f : \mathcal{X} \to \mathcal{G}_m$ of the minimizer f^* of the expected risk $\mathcal{R}(f) = \mathbb{E}_\rho[\mathrm{FNGW}_{\alpha,\beta}(f(X), G)]$

Proposed Estimator

Based on the work of [Ciliberto et al., 2020, Brogat-Motte et al., 2022], we propose an estimator of the following form

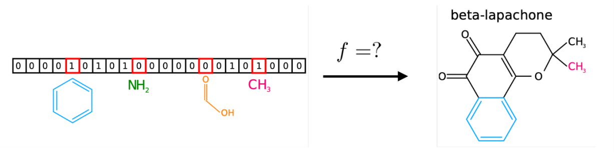
$$\hat{f}(x) = \underset{g \in \mathcal{G}_m}{\arg\min} \sum_{i=1}^{n} \xi(x)_i \mathrm{FNGW}_{\alpha,\beta}(g, g_i) \tag{7}$$

with $\xi(x) = \mathrm{K}S^\top(S\mathrm{K}^2S^\top + n\lambda S\mathrm{K}S^\top)^\dagger S\boldsymbol{\kappa}_x$ where $\mathrm{K} \in \mathbb{R}^{n \times n}$ is the input kernel Gram matrix, $\boldsymbol{\kappa}_x = (k(x, x_1), \ldots, k(x, x_n))^\top \in \mathbb{R}^n$, and $S \in \mathbb{R}^{s \times n}$ with $s \ll n$ is a sketching matrix.

- (Proposition) The FNGW loss admits an Implicit Loss Embedding (ILE) $\rightarrow \hat{f}$ is universally consistent and its learning rate is of order $n^{-1/4}$ with additional assumptions.
- Sketched ILE enables the supervised graph prediction with more than 100,000 training data points.

The estimator describes actually a barycenter problem.

Experiment: Fingerprint to Molecule



Fin2Mol Dataset:

- Predict a QM9 molecule from its fingerprint.
- Each molecule contains up to 9 atoms.
- The dataset contains around 130,000 fingerprint-molecule pairs.

Table 1: Graph edit distances of different methods on the Fin2Mol test set.

	GED w/o edge feature ↓	GED w/ edge feature ↓
NNBary-FGW	5.000 ± 0.140	-
NNBary-FNGW	5.311 ± 0.090	5.756 ± 0.073
Sketched ILE-FGW	3.037 ± 0.111	-
Sketched ILE-FNGW	$\mathbf{1.449 \pm 0.034}$	$\mathbf{1.534 \pm 0.029}$

Experiment: Fingerprint to Molecule

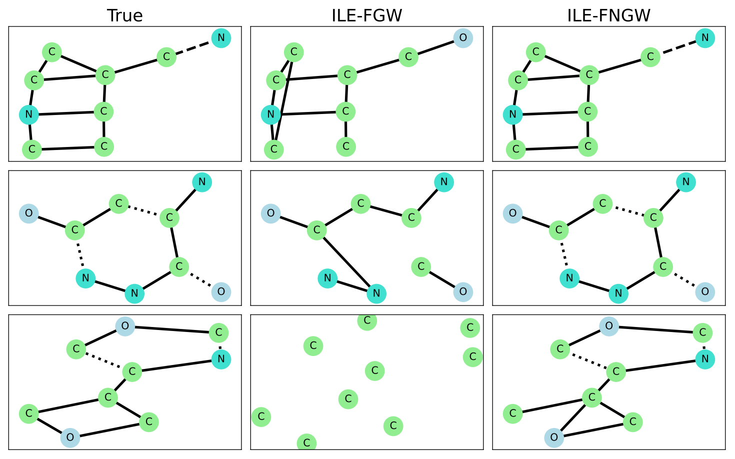


Figure 5: Qualitative comparison of the predicted QM9 molecules.

Experiment: Metabolite Identification

Table 2: Top-k accuracies on the metabolite identification test set. Best results are in **Bold**.

	Top-1 ↑	Top-10 ↑	Top-20 ↑
WL kernel	9.8%	29.1%	37.4%
IOKR - Fingerprint w/ linear kernel	28.6%	54.5%	59.9%
IOKR - Fingerprint w/ gaussian kernel	**41.0%**	**62.0%**	**67.8%**
ILE-FGW diffuse	28.1%	53.6%	59.9%
ILE-FNGW diffuse + Bond stereo	27.7%	55.2%	60.9%
ILE-FNGW diffuse + Bond type	34.6%	55.1%	60.0%
ILE-FNGW diffuse + Mix	36.2%	58.2%	61.9%

Conclusion and Future Work

- FNGW inherits similar geometric properties as FGW and NGW.
- FNGW benefits supervised graph prediction.
- Acceleration of both the distance computation and the barycenter computation.
- Integration of our codes into POT[2] package.
- Potential usage of FNGW in other graph learning algorithms where the pairwise graph comparison is involved.

[2] POT: Python Optimal Transport, `https://pythonot.github.io/`

Thanks for your attention!

Questions?

Brogat-Motte, L., Flamary, R., Brouard, C., Rousu, J., and d'Alché Buc, F. (2022).
Learning to Predict Graphs with Fused Gromov-Wasserstein Barycenters.
In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2321–2335. PMLR.

Brouard, C., Shen, H., Dührkop, K., d'Alché Buc, F., Böcker, S., and Rousu, J. (2016).
Fast Metabolite Identification with Input Output Kernel Regression.
Bioinformatics, 32(12):i28–i36.

References ii

Chowdhury, S. and Mémoli, F. (2019).
The Gromov–Wasserstein Distance Between Networks and Stable Network Invariants.
Information and Inference: A Journal of the IMA, 8(4):757–787.

Ciliberto, C., Rosasco, L., and Rudi, A. (2020).
A General Framework for Consistent Structured Prediction with Implicit Loss Embeddings.
Journal of Machine Learning Research, 21(98):1–67.

Mémoli, F. (2011).
Gromov–Wasserstein Distances and the Metric Approach to Object Matching.
Foundations of Computational Mathematics, 11(4):417–487.

Peyré, G., Cuturi, M., and Solomon, J. (2016).
Gromov-Wasserstein Averaging of Kernel and Distance Matrices.
In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2664–2672. PMLR.

Sturm, K.-T. (2012).
The Space of Spaces: Curvature Bounds and Gradient Flows on the Space of Metric Measure Spaces.
arXiv preprint arXiv:1208.0434.

Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. (2019).
Optimal Transport for structured data with application on graphs.
In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6275–6284. PMLR.

Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. (2020).
Fused Gromov-Wasserstein Distance for Structured Objects.
Algorithms, 13(9):212.

Vincent-Cuaz, C., Vayer, T., Flamary, R., Corneli, M., and Courty, N. (2021).
Online Graph Dictionary Learning.
In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10564–10574. PMLR.

FNGW Distance, General Form

(Definition): FNGW Distance, General Form

Let $\mathcal{G}$ be the set of tuples of the form $(X, \psi_X, \varphi_X, \omega_X, \mu_X)$ where X is a polish space, $\psi_X : X \to \Psi$ is a bounded continuous measurable function from X to a metric space (Ψ, d_Ψ) , $\varphi_X : X \times X \to \mathbb{R}$ is a bounded continuous measurable function, $\omega_X : X \times X \to \Omega$ is a bounded continuous measurable function from X^2 to a metric space (Ω, d_Ω) and μ_X is a fully supported Borel probability measure.

(Definition): FNGW Distance, General Form

Given two tuples $g_X = (X, \psi_X, \varphi_X, \omega_X, \mu_X)$, $g_Y = (Y, \psi_Y, \varphi_Y, \omega_Y, \mu_Y)$ from $\mathcal{G}$ and trade-off parameters $(\alpha, \beta) \in [0,1]^2$, the Fused Network Gromov-Wasserstein Distance between g_X and g_Y is defined for any $(p, q) \in [1, \infty]$ as follows:

$$\mathrm{FNGW}_{\alpha,\beta,q,p}(g_X, g_Y) = \min_{\mu \in \Pi(\mu_X, \mu_Y)} \mathcal{E}_{\alpha,\beta,q,p}(g_X, g_Y, \mu) \tag{8}$$

with

$$\mathcal{E}_{\alpha,\beta,q,p}(g_X, g_Y, \mu) = \Bigg(\int_{X\times Y} \int_{X\times Y} [(1-\alpha-\beta) d_\Psi (\psi_X(x), \psi_Y(y))^q$$
$$+ \alpha d_\Omega(\omega_X(x,x'), \omega_Y(y,y'))^q + \beta |\varphi_X(x,x') - \varphi_Y(y,y')|^q]^p d\mu(x,y) d\mu(x',y') \Bigg)^{\frac{1}{p}} \tag{9}$$

Metric Properties

We verify the metric properties satisfied by FNGW distance in the **general case**.

Theorem (Metric Properties)

The FNGW distance satisfies the following properties: for all $g_X = (X, \psi_X, \varphi_X, \omega_X, \mu_X)$, $g_Y = (Y, \psi_Y, \varphi_Y, \omega_Y, \mu_Y)$ and $g_Z = (Z, \psi_Z, \varphi_Z, \omega_Z, \mu_Z)$ from $\mathcal{G}$:

- (Positivity) $\mathrm{FNGW}_{\alpha,\beta,q,p}(g_X, g_Y) \geq 0$
- (Symmetry) $\mathrm{FNGW}_{\alpha,\beta,q,p}(g_X, g_Y) = \mathrm{FNGW}_{\alpha,\beta,q,p}(g_Y, g_X)$
- (Equality) $\mathrm{FNGW}_{\alpha,\beta,q,p}(g_X, g_X) = 0$. $\mathrm{FNGW}_{\alpha,\beta,q,p}(g_X, g_Y) = 0$ if and only if g_X is weakly isomorphic to g_Y.
- (Relaxed Triangle Inequality) $\mathrm{FNGW}_{\alpha,\beta,q,p}(g_X, g_Z) \leq 2^{q-1}(\mathrm{FNGW}_{\alpha,\beta,q,p}(g_X, g_Y) + \mathrm{FNGW}_{\alpha,\beta,q,p}(g_Y, g_Z))$

Weak Isomorphism

(Definition) Weak Isomorphism of Node and Edge Featured Graphs

Two graphs g_X and g_Y are isomorphic if and only there is a Borel probability space (Z, μ_Z) with measurable maps $f : Z \to X$ and $g : Z \to Y$ such that

$$f_{\#}\mu_Z = \mu_X \quad g_{\#}\mu_Z = \mu_Y \tag{10}$$

$$\begin{aligned} \|(1-\alpha-\beta)d_{\Psi}\left(\psi_X \circ f, \psi_Y \circ g\right)^q + \alpha d_{\Omega}(f^{\#}\omega_X, g^{\#}\omega_Y)^q \\ +\beta|f^{\#}\varphi_X - g^{\#}\varphi_Y|^q\|_\infty = 0 \end{aligned} \tag{11}$$

FNGW Computation Algorithm

Algorithm 1 Computation of the FNGW Distance by CGD

input: $g = (F, A, E, \boldsymbol{p})$, $\tilde{g} = (\tilde{F}, \tilde{A}, \tilde{E}, \tilde{\boldsymbol{p}})$ and trade-off parameters (α, β)
init: $\pi^{(0)} = \boldsymbol{p}\tilde{\boldsymbol{p}}^\mathsf{T} \in \mathbb{R}^{m \times \tilde{m}}$
for $k = 1, \dots, K$ **do**
Calculate gradient: $G = \nabla_{\pi^{(k-1)}} \mathcal{E}_{\alpha,\beta}((F, A, E), (\tilde{F}, \tilde{A}, \tilde{E}), \pi^{(k-1)})$
Solve the optimization problem with an OT solver: $\tilde{\pi}^{(k-1)} \in \arg\min_{\tilde{\pi} \in \Pi(\boldsymbol{p}, \tilde{\boldsymbol{p}})} \langle G, \tilde{\pi} \rangle$
Update the optimal plan: $\pi^{(k)} = (1 - \gamma^{(k)})\pi^{(k-1)} + \gamma^{(k)}\tilde{\pi}^{(k-1)}$ with $\gamma^{(k)} \in (0, 1)$ given by line-search algorithm (See details in Appendix B).
end for
Calculate the distance: $\mathrm{FNGW}_{\alpha,\beta}(g, \tilde{g}) = \mathcal{E}_{\alpha,\beta}((F, A, E), (\tilde{F}, \tilde{A}, \tilde{E}), \pi^{(K)})$
output: $\mathrm{FNGW}_{\alpha,\beta}(g, \tilde{g})$ and $\pi^{(K)}$

FNGW Barycenter Algorithm

Algorithm 2 Computation of FNGW Barycenter with BCD

input: $\{g_i\}_i$, fixed histogram $\boldsymbol{p}$, trade-off parameter (α, β)
init: Randomly initialize $E^{(0)}, F^{(0)}$ and $A^{(0)}$.
for $k = 1, \ldots, K$ **do**
  Calculate $\{\pi_i\}_i$ with Alg. 1: $\pi_i^{(k)} = \arg\min_{\pi_i \in \Pi(\boldsymbol{p}, \boldsymbol{p}_i)} \mathcal{E}_{\alpha,\beta}\left((F^{(k-1)}, A^{(k-1)}, E^{(k-1)}), (F_i, A_i, E_i), \pi_i\right)$
  Update E: $E^{(k)} = \frac{1}{\mathcal{I}_{m\times T} \times_2 \boldsymbol{p}\boldsymbol{p}^\top} \sum_i \lambda_i (E_i \times_2 \pi_i^{(k)}) \times_1 \pi_i^{(k)}$ ▷ Prop. 2.10
  Update A: $A^{(k)} = \frac{1}{\boldsymbol{p}\boldsymbol{p}^\top} \sum_i \lambda^i \pi_i^{(k)} A_i \pi_i^{(k)\top}$ ▷ Prop. 4 in Peyré et al. (2016)
  Update F: $F^{(k)} = \sum_i \lambda_i \mathrm{diag}(\frac{1}{\boldsymbol{p}}) \pi_i^{(k)} F_i$ ▷ Equation 8 in Cuturi & Doucet (2014)
end for
output: The barycenter $(F^{(K)}, A^{(K)}, E^{(K)})$

ILE Condition

(Definition): Implicit Loss Embedding

A loss function $\Delta : \mathcal{Y} \times \mathcal{Y}$ is said to admit an Implicit Loss Embedding (ILE) if there exist a separable Hilbert space $\mathcal{Z}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{Z}}$, a continuous embedding $\psi : \mathcal{Y} \to \mathcal{Z}$ and a bounded linear operator $V : \mathcal{Z} \to \mathcal{Z}$ such that for all $y, y' \in \mathcal{Y}$

$$\Delta(y, y') = \langle \psi(y), V\psi(y') \rangle_{\mathcal{Z}} \tag{12}$$