Supplementary material for Online Graph Dictionary Learning

1. Notations & definitions

In this section we recall the notations used in the rest of the supplementary.

For matrices we note $S_N(\mathbb{R})$ the set of symmetric matrices in $\mathbb{R}^{N\times N}$ and $\langle\cdot,\cdot\rangle_F$ the Frobenius inner product defined for real matrices $\boldsymbol{C}_1, \boldsymbol{C}_2$ as $\langle \boldsymbol{C}_1, \boldsymbol{C}_2\rangle_F = \text{tr}(\boldsymbol{C}_1^\top \boldsymbol{C}_2)$ where tr denotes the trace of matrices. Moreover $\boldsymbol{C}_1 \odot \boldsymbol{C}_2$ denotes the Hadamard product of $\boldsymbol{C}_1, \boldsymbol{C}_2$, *i.e.* $(\boldsymbol{C}_1 \odot \boldsymbol{C}_2)_{ij} = C_1(i,j)C_2(i,j)$. Finally $\text{vec}(\boldsymbol{C})$ denotes the vectorization of the matrix $\boldsymbol{C}$.

For vectors the Euclidean norm is denoted as $\|\cdot\|_2$ associated with the inner product $\langle\cdot,\cdot\rangle$. For a vector $\mathbf{x}\in\mathbb{R}^N$ the operator $\text{diag}(\mathbf{x})$ denotes the diagonal matrix defined with the values of $\mathbf{x}$. If $\mathbf{M}\in S_N(\mathbb{R})$ is a positive semi-definite matrix we note $\|\cdot\|_{\mathbf{M}}$ the pseudo-norm defined for $\mathbf{x}\in\mathbb{R}^N$ by $\|\mathbf{x}\|^2_{\mathbf{M}} = \mathbf{x}^\top\mathbf{M}\mathbf{x}$. By some abuse of terminology we will use the term Mahalanobis distance to refer to generalized quadratic distances defined as $d_{\mathbf{M}}(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_{\mathbf{M}}$. The fact that $\mathbf{M}$ is positive semi-definite ensures that $d_{\mathbf{M}}$ satisfies the properties of a pseudo-distance.

For a 4-D tensor $\mathbf{L} = (L_{ijkl})_{ijkl}$ we note $\otimes$ the tensor-matrix multiplication, *i.e.* given a matrix $\boldsymbol{C}$, $\mathbf{L}\otimes\mathbf{A}$ is the matrix $\left(\sum_{k,l} L_{i,j,k,l}A_{k,l}\right)_{i,j}$.

The simplex of histograms (or *weights*) with N bins is $\Sigma_N := \left\{\mathbf{h}\in\mathbb{R}_N^+ | \sum_i h_i = 1\right\}$. For two histograms $\boldsymbol{h}^X \in \Sigma_{N_X}, \boldsymbol{h}^Y\in\Sigma_{N_Y}$ the set $\mathcal{U}(\boldsymbol{h}^X,\boldsymbol{h}^Y) := \{\boldsymbol{T}\in\mathbb{R}_+^{N^X\times N^Y} | \boldsymbol{T}\mathbf{1}_{N^Y} = \boldsymbol{h}^X, \boldsymbol{T}^T\mathbf{1}_{N^X} = \boldsymbol{h}^Y\}$ is the set of couplings between $\boldsymbol{h}^X, \boldsymbol{h}^Y$.

Recall that for two graphs $G^X = (\boldsymbol{C}^X, \mathbf{h}^X)$ and $G^Y = (\boldsymbol{C}^Y,\mathbf{h}^Y)$ the GW_2 distance between G^X and G^Y is defined as the result of the following optimization problem:

$$\min_{\boldsymbol{T}\in\mathcal{U}(\boldsymbol{h}^X,\boldsymbol{h}^Y)} \sum_{ijkl}\left(C^X_{ij} - C^Y_{kl}\right)^2 T_{ik}T_{jl} \tag{1}$$

In the following we denote by $GW_2(\boldsymbol{C}^X,\boldsymbol{C}^Y,\boldsymbol{h}^X,\boldsymbol{h}^Y)$ the optimal value of equation 1 or by $GW_2(\boldsymbol{C}^X,\boldsymbol{C}^Y)$ when the weights are uniform. With more compact notations:

$$GW_2(\boldsymbol{C}^X,\boldsymbol{C}^Y,\boldsymbol{h}^X,\boldsymbol{h}^Y) = \min_{\boldsymbol{T}\in\mathcal{U}(\boldsymbol{h}^X,\boldsymbol{h}^Y)} \langle \mathbf{L}(\boldsymbol{C}^X,\boldsymbol{C}^Y)\otimes\boldsymbol{T},\boldsymbol{T}\rangle_F \tag{2}$$

where $\mathbf{L}(\boldsymbol{C}^X,\boldsymbol{C}^Y)$ is the 4-D tensor $\mathbf{L}(\boldsymbol{C}^X,\boldsymbol{C}^Y) = \left((C^X_{ij} - C^Y_{kl})^2\right)_{ijkl}$

For graphs with attributes we use the Fused Gromov-Wasserstein distance (Vayer et al., 2019). More precisely consider two graphs $G^X = (\boldsymbol{C}^X,\boldsymbol{A}^X,\mathbf{h}^X)$ and $G^Y = (\boldsymbol{C}^Y,\boldsymbol{A}^Y,\mathbf{h}^Y)$ where $\boldsymbol{A}^X = (\mathbf{a}^X_i)_{i\in[N^X]}\in\mathbb{R}^{N^X\times d}$, $\boldsymbol{A}^Y = (\mathbf{a}^Y_j)_{j\in[N^Y]}\in\mathbb{R}^{N^Y\times d}$ are the matrices of all features. Given $\alpha\in[0,1]$ and a cost function $c:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ between vectors in $\mathbb{R}^d$ the FGW_2 distance is defined as the result of the following optimization problem:

$$\min_{\boldsymbol{T}\in\mathcal{U}(\boldsymbol{h}^X,\boldsymbol{h}^Y)} (1-\alpha)\sum_{ij} c(\mathbf{a}^X_i,\mathbf{a}^Y_j)T_{ij} + \alpha\sum_{ijkl}\left(C^X_{ij} - C^Y_{kl}\right)^2 T_{ik}T_{jl} \tag{3}$$

In the following we note $FGW_{2,\alpha}(\boldsymbol{C}^X,\boldsymbol{A}^X,\boldsymbol{C}^Y,\boldsymbol{A}^Y,\boldsymbol{h}^X,\boldsymbol{h}^Y)$ the optimal value of equation 3 or by $FGW_{2,\alpha}(\boldsymbol{C}^X,\boldsymbol{A}^X,\boldsymbol{C}^Y,\boldsymbol{A}^Y)$ when the weights are uniform. The term $\sum_{ij} c(\mathbf{a}^X_i,\mathbf{a}^Y_j)T_{ij}$ will be called the *Wasserstein objective* and denoted as $\mathcal{F}(\boldsymbol{A}^X,\boldsymbol{A}^Y,\boldsymbol{T})$ and the term $\sum_{ijkl}\left(C^X_{ij} - C^Y_{kl}\right)^2 T_{ik}T_{jl}$ will be called the *Gromov-Wasserstein objective* and denoted $\mathcal{E}(\boldsymbol{C}^X,\boldsymbol{C}^Y,\boldsymbol{T})$.

2. Proofs of the different results

2.1. (F)GW upper-bounds in the embedding space

Proposition 1 (Gromov-Wasserstein) *For two embedded graphs with embeddings $\boldsymbol{w}^{(1)}$ and $\boldsymbol{w}^{(2)}$ over the set of pairwise relation matrices $\{\overline{\boldsymbol{C}_s}\}_{s\in[S]} \subset S_N(\mathbb{R})$, with a shared masses vector $\boldsymbol{h}$, the following inequality holds*

$$GW_2\left(\sum_{s\in[S]} w_s^{(1)}\overline{\boldsymbol{C}_s}, \sum_{s\in[S]} w_s^{(2)}\overline{\boldsymbol{C}_s}\right) \leq \|\boldsymbol{w}^{(1)} - \boldsymbol{w}^{(2)}\|_{\boldsymbol{M}} \tag{4}$$

where $\mathrm{M} = (\langle \boldsymbol{D_h}\overline{\boldsymbol{C}_p}, \overline{\boldsymbol{C}_q}\boldsymbol{D_h}\rangle_F)_{pq}$ and $\boldsymbol{D_h} = diag(\boldsymbol{h})$. $\boldsymbol{M}$ is a positive semi-definite matrix hence engenders a Mahalanobis distance between embeddings.

Proof. Let consider the formulation of the GW distance as a Frobenius inner product (see *e.g* (Peyré et al., 2016)). Denoting $\boldsymbol{T}$ the optimal transport plan between both embedded graph and the power operation over matrices applied at entries level,

$$GW_2(\sum_s w_s^{(1)}\overline{\boldsymbol{C}_s}, \sum_s w_s^{(2)}\overline{\boldsymbol{C}_s}, \boldsymbol{h}) = \langle (\sum_s w_s^{(1)}\overline{\boldsymbol{C}_s})^2\boldsymbol{h}\boldsymbol{1}_N^\top + \boldsymbol{1}_N\boldsymbol{h}^\top(\sum_s w_s^{(2)}\overline{\boldsymbol{C}_s}^\top)^2 - 2(\sum_s w_s^{(1)}\overline{\boldsymbol{C}_s})\boldsymbol{T}(\sum_s w_s^{(2)}\overline{\boldsymbol{C}_s}^\top), \boldsymbol{T}\rangle_F \tag{5}$$

Using the marginal constraints of GW problem, *i.e* $\boldsymbol{T} \in \mathcal{U}(\boldsymbol{h},\boldsymbol{h}) := \{\boldsymbol{T} \in \mathbb{R}_+^{N\times N} | \boldsymbol{T}\boldsymbol{1}_N = \boldsymbol{h}, \boldsymbol{T}^T\boldsymbol{1}_N = \boldsymbol{h}\}$, and the symmetry of matrices $\{\overline{\boldsymbol{C}_s}\}$,equation 5 can be developed as follow,

$$GW_2(\sum_s w_s^{(1)}\overline{\boldsymbol{C}_s}, \sum_s w_s^{(2)}\overline{\boldsymbol{C}_s}, \boldsymbol{h}) = \sum_{pq} \operatorname{tr}\left(w_p^{(1)}w_q^{(1)}(\overline{\boldsymbol{C}_p}\odot\overline{\boldsymbol{C}_q})\boldsymbol{h}\boldsymbol{h}^\top + w_p^{(2)}w_q^{(2)}(\overline{\boldsymbol{C}_p}\odot\overline{\boldsymbol{C}_q})\boldsymbol{h}\boldsymbol{h}^\top - 2w_p^{(1)}w_q^{(2)}\overline{\boldsymbol{C}_p}\boldsymbol{T}\overline{\boldsymbol{C}_q}\boldsymbol{T}^\top\right) \tag{6}$$

With the following property of the trace operator:

$$\operatorname{tr}\left((\boldsymbol{C}_1 \odot \boldsymbol{C}_2)\boldsymbol{x}\boldsymbol{x}^\top\right) = \operatorname{tr}\left(\boldsymbol{C}_1^\top \operatorname{diag}(\boldsymbol{x})\boldsymbol{C}_2\operatorname{diag}(\boldsymbol{x})\right) \tag{7}$$

Denoting $\boldsymbol{D_h} = \operatorname{diag}(\boldsymbol{h})$, equation 6 can be expressed as:

$$GW_2(\sum_p w_p^{(1)}\overline{\boldsymbol{C}_p}, \sum_q w_q^{(2)}\overline{\boldsymbol{C}_q}, \boldsymbol{h}) = \sum_{pq}(w_p^{(1)}w_q^{(1)} + w_p^{(2)}w_q^{(2)})\langle \boldsymbol{D_h}\overline{\boldsymbol{C}_p}, \overline{\boldsymbol{C}_q}\boldsymbol{D_h}\rangle_F - 2w_p^{(1)}w_q^{(2)}\langle \boldsymbol{T}^\top\overline{\boldsymbol{C}_p}, \overline{\boldsymbol{C}_q}\boldsymbol{T}^\top\rangle_F \tag{8}$$

As $\boldsymbol{T} \in \mathcal{U}(\boldsymbol{h},\boldsymbol{h})$ is a minimum of the GW objective, we can bound by above equation 6 by evaluating the GW objective in $\boldsymbol{D_h} \in \mathcal{U}(\boldsymbol{h},\boldsymbol{h})$, which is a sub-optimal admissible coupling.

$$\begin{aligned} GW_2(\sum_p w_p^{(1)}\overline{\boldsymbol{C}_p}, \sum_q w_q^{(2)}\overline{\boldsymbol{C}_q}, \boldsymbol{h}) &\leq \sum_{pq}(w_p^{(1)}w_q^{(1)} + w_p^{(2)}w_q^{(2)} - 2w_p^{(1)}w_q^{(2)})\langle \boldsymbol{D_h}\overline{\boldsymbol{C}_p}, \overline{\boldsymbol{C}_q}\boldsymbol{D_h}\rangle_F \\ &= {\boldsymbol{w}^{(1)}}^T\boldsymbol{M}\boldsymbol{w}^{(1)} + {\boldsymbol{w}^{(2)}}^\top\boldsymbol{M}\boldsymbol{w}^{(2)} - 2{\boldsymbol{w}^{(1)}}^\top\boldsymbol{M}\boldsymbol{w}^{(2)} \end{aligned} \tag{9}$$

with $\boldsymbol{M} = (\langle \boldsymbol{D_h}\overline{\boldsymbol{C}_p}, \overline{\boldsymbol{C}_q}\boldsymbol{D_h}\rangle_F)_{pq}$. It suffices to prove that the matrix $\boldsymbol{M}$ is a PSD matrix to conclude that it defines a Mahalanobis distance over the set of embeddings $\boldsymbol{w}$ which bounds by above the GW distance between corresponding embedded graphs. Let consider the following reformulation of an entry M_{pq} as follow,

$$\langle \boldsymbol{D_h}\overline{\boldsymbol{C}_p}, \overline{\boldsymbol{C}_q}\boldsymbol{D_h}\rangle = \operatorname{vec}(\boldsymbol{B}_p)^\top\operatorname{vec}(\boldsymbol{B}_q) \tag{10}$$

where $\forall n \in [S], \boldsymbol{B}_n = \boldsymbol{D_h}^{1/2}\overline{\boldsymbol{C}_n}\boldsymbol{D_h}^{1/2}$. Hence with $\boldsymbol{B} = (\boldsymbol{B}_n)_n \subset \mathbb{R}^{N^2\times S}$, $\boldsymbol{M}$ can be factorized as $\boldsymbol{B}^T\boldsymbol{B}$ and therefore is a PSD matrix. □

A similar result can be proven for the Fused Gromov-Wasserstein distance:

Proposition 2 (Fused Gromov-Wasserstein) *For two embedded graphs with node attributes, with embeddings $\boldsymbol{w}^{(1)}$ and $\boldsymbol{w}^{(2)}$ over the set of pairwise relation matrices $\{(\overline{\boldsymbol{C}_s}, \overline{\boldsymbol{A}_s})\}_{s\in[S]} \subset S_N(\mathbb{R}) \times \mathbb{R}^{N\times dd}$, and a shared masses vector $\boldsymbol{h}$, the following inequality holds $\forall\alpha \in (0,1)$,*

$$FGW_{2,\alpha}\left(\widetilde{\boldsymbol{C}}(\boldsymbol{w}^{(1)}), \widetilde{\boldsymbol{A}}(\boldsymbol{w}^{(1)}), \widetilde{\boldsymbol{C}}(\boldsymbol{w}^{(2)}), \widetilde{\boldsymbol{A}}(\boldsymbol{w}^{(2)})\right) \leq \|\boldsymbol{w}^{(1)} - \boldsymbol{w}^{(2)}\|_{\alpha\boldsymbol{M}_1 + (1-\alpha)\boldsymbol{M}_2} \tag{11}$$

with,

$$\widetilde{\boldsymbol{C}}(\boldsymbol{w}) = \sum_s w_s \overline{\boldsymbol{C}_s} \quad \textit{and} \quad \widetilde{\boldsymbol{A}}(\boldsymbol{w}) = \sum_s w_s \overline{\boldsymbol{A}_s} \tag{12}$$

Where $\boldsymbol{M}_1 = \left(\langle \boldsymbol{D_h} \overline{\boldsymbol{C}_p}, \overline{\boldsymbol{C}_q} \boldsymbol{D_h} \rangle_F \right)_{pq}$ *and* $\boldsymbol{M}_2 = (\langle \boldsymbol{D}_{\boldsymbol{h}}^{1/2} \overline{\boldsymbol{A}_p}, \boldsymbol{D}_{\boldsymbol{h}}^{1/2} \overline{\boldsymbol{A}_q} \rangle_F)_{pq \in [S]}$*, and* $\boldsymbol{D_h} = diag(\boldsymbol{h})$*, are PSD matrices and therefore their linear combination being PSD engender Mahalanobis distances over the unmixing space.*

Proof. Let consider the optimal transport plan $\boldsymbol{T} \in \mathcal{U}(\boldsymbol{h}, \boldsymbol{h})$ of the FGW distance between both embedded structures.

$$FGW_{2,\alpha}^2 \left(\widetilde{\boldsymbol{C}}(\boldsymbol{w}^{(1)}), \widetilde{\boldsymbol{A}}(\boldsymbol{w}^{(1)}), \widetilde{\boldsymbol{C}}(\boldsymbol{w}^{(2)}), \widetilde{\boldsymbol{A}}(\boldsymbol{w}^{(2)}), \boldsymbol{h} \right) = \alpha \mathcal{E} \left(\widetilde{\boldsymbol{C}}(\boldsymbol{w}^{(1)}), \widetilde{\boldsymbol{C}}(\boldsymbol{w}^{(2)}), \boldsymbol{T} \right) + (1-\alpha) \mathcal{F} \left(\widetilde{\boldsymbol{A}}(\boldsymbol{w}^{(1)}), \widetilde{\boldsymbol{A}}(\boldsymbol{w}^{(2)}), \boldsymbol{T} \right) \tag{13}$$

where $\mathcal{E}$ and $\mathcal{F}$ denotes respectively the Gromov-Wasserstein objective and the Wasserstein objective. As a similar approach than for Proposition 4 can be used for the GW objective involved in equation 13, we will first highlight a suitable factorization of the Wasserstein objective $\mathcal{F}$. Note that for any feature matrices $\boldsymbol{A}_1 = (\boldsymbol{a}_{1,i})_{i \in [N]}, \boldsymbol{A}_2 = (\boldsymbol{a}_{2,i})_{i \in [N]} \in \mathbb{R}^{N * d}$, $\mathcal{F}$ with an euclidean ground cost can be expressed as follow using the marginal constraints on $\boldsymbol{T} \in \mathcal{U}(\boldsymbol{h}, \boldsymbol{h})$,

$$\begin{aligned}
\mathcal{F}(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{T}) &= \sum_{ij} \|\boldsymbol{a}_{1,i} - \boldsymbol{a}_{2,j}\|_2^2 T_{ij} \\
&= \sum_i \|\boldsymbol{a}_{1,i}\|_2^2 h_i + \sum_j \|\boldsymbol{a}_{1,j}\|_2^2 h_j - 2 \sum_{ij} \langle \boldsymbol{a}_{1,i}, \boldsymbol{a}_{2,j} \rangle T_{ij} \\
&= \langle \boldsymbol{D}_{\boldsymbol{h}}^{1/2} \boldsymbol{A}_1, \boldsymbol{D}_{\boldsymbol{h}}^{1/2} \boldsymbol{A}_1 \rangle_F + \langle \boldsymbol{D}_{\boldsymbol{h}}^{1/2} \boldsymbol{A}_2, \boldsymbol{D}_{\boldsymbol{h}}^{1/2} \boldsymbol{A}_2 \rangle_F - 2 \langle \boldsymbol{A}_1 \boldsymbol{A}_2^\top, \boldsymbol{T} \rangle_F
\end{aligned} \tag{14}$$

Returning to our main problem 13, a straigth-forward development of its Wasserstein term $\mathcal{F}$ using equation 14 leads to the following equality,

$$\mathcal{F}\left(\widetilde{\boldsymbol{A}}(\boldsymbol{w}^{(1)}), \widetilde{\boldsymbol{A}}(\boldsymbol{w}^{(2)}), \boldsymbol{T} \right) = \sum_{pq} \left(w_p^{(1)} w_q^{(1)} + w_p^{(2)} w_q^{(2)} \right) \langle \boldsymbol{D}_{\boldsymbol{h}}^{1/2} \overline{\boldsymbol{A}_p}, \boldsymbol{D}_{\boldsymbol{h}}^{1/2} \overline{\boldsymbol{A}_q} \rangle_F - 2 w_p^{(1)} w_q^{(2)} \langle \boldsymbol{A}_p \boldsymbol{A}_q^\top, \boldsymbol{T} \rangle_F \tag{15}$$

Similarly than for the proof of Proposition 1, $\boldsymbol{T} \in \mathcal{U}(\boldsymbol{h}, \boldsymbol{h})$ is an optimal admissible coupling minimizing the FGW problem, thus equation 13 is upper bounded by its evaluation in the sub-optimal admissible coupling $\boldsymbol{D_h} \in \mathbf{U}(\boldsymbol{h}, \boldsymbol{h})$. Let $\boldsymbol{M}_1 = \boldsymbol{M} = (\langle \boldsymbol{D_h} \overline{\boldsymbol{C}_p}, \overline{\boldsymbol{C}_q} \boldsymbol{D_h} \rangle_F)_{pq}$ the PSD matrix coming from the proof of Proposition 1.

Let $\boldsymbol{M}_2 = \left(\langle \boldsymbol{D}_{\boldsymbol{h}}^{1/2} A_p, \boldsymbol{D}_{\boldsymbol{h}}^{1/2} A_q \rangle_F \right)_{pq}$ which is also a PSD matrix as it can be factorized as $\boldsymbol{B}^\top \boldsymbol{B}$ with $\boldsymbol{B} = \left(vec(\boldsymbol{D}_{\boldsymbol{h}}^{1/2} \boldsymbol{A}_s) \right)_{s \in [S]} \in \mathbb{R}^{Nd \times S}$.

Let us denote $\forall \alpha \in (0,1)$, $\boldsymbol{M}_\alpha = \alpha \boldsymbol{M}_1 + (1-\alpha) \boldsymbol{M}_2$ which is PSD as convex combination of PSD matrices, hence engender a Mahalanobis distance in the embedding space. To summarize, equation 16 holds $\forall \alpha \in (0,1)$,

$$\begin{aligned}
FGW_{2,\alpha}^2 \left(\widetilde{\boldsymbol{C}}(\boldsymbol{w}^{(1)}), \widetilde{\boldsymbol{A}}(\boldsymbol{w}^{(1)}), \widetilde{\boldsymbol{C}}(\boldsymbol{w}^{(2)}), \widetilde{\boldsymbol{A}}(\boldsymbol{w}^{(2)}), \boldsymbol{h} \right) &\leq {\boldsymbol{w}^{(1)}}^\top \boldsymbol{M}_\alpha \boldsymbol{w}^{(1)} + {\boldsymbol{w}^{(2)}}^\top \boldsymbol{M}_\alpha \boldsymbol{w}^{(2)} - 2 {\boldsymbol{w}^{(1)}}^\top \boldsymbol{M}_\alpha \boldsymbol{w}^{(2)} \\
&= \|\boldsymbol{w}^{(1)} - \boldsymbol{w}^{(2)}\|_{\boldsymbol{M}_\alpha}
\end{aligned} \quad \square \tag{16}$$

2.2. Proposition 3. Gradients of GW *w.r.t.* the weights

In this section we will prove the following result:

Proposition 3 *Let* $(\boldsymbol{C}^1, \boldsymbol{h}^1)$ *and* $(\boldsymbol{C}^2, \boldsymbol{h}^2)$ *be two graphs. Let* $\boldsymbol{T}^*$ *be an optimal coupling of the GW problem between* $(\boldsymbol{C}^1, \boldsymbol{h}^1), (\boldsymbol{C}^2, \boldsymbol{h}^2)$*. We define the following cost matrix* $\boldsymbol{M}(\boldsymbol{T}^*) := \left(\sum_{kl} (C^1_{ik} - C^2_{jl})^2 T^*_{kl} \right)_{ij}$*. Let* $\boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{\beta}^*(\boldsymbol{T}^*)$ *be the dual variables of the following linear OT problem:*

$$\min_{\boldsymbol{T} \in \mathcal{U}(\boldsymbol{h}^1, \boldsymbol{h}^2)} \langle \boldsymbol{M}(\boldsymbol{T}^*), \boldsymbol{T} \rangle_F$$

Then $\boldsymbol{\alpha}^*(\boldsymbol{T}^*)$ *(resp* $\boldsymbol{\beta}^*(\boldsymbol{T}^*)$*) is a subgradient of the function* $GW_2^2(\boldsymbol{C}^1, \boldsymbol{C}^2, \bullet, \boldsymbol{h}^2)$ *(resp* $GW_2^2(\boldsymbol{C}^1, \boldsymbol{C}^2, \boldsymbol{h}^1, \bullet)$*).*

In the following $\boldsymbol{T} \geq 0$ should be understood as $\forall i, j\ T_{ij} \geq 0$. Let $(\boldsymbol{C}^1, \boldsymbol{h}^1)$ and $(\boldsymbol{C}^2, \boldsymbol{h}^2)$ be two graphs of order n and m with $\boldsymbol{C}^1 \in S_n(\mathbb{R}), \boldsymbol{C}^2 \in S_m(\mathbb{R})$ and $(\boldsymbol{h}^1, \boldsymbol{h}^2) \in \Sigma_n \times \Sigma_m$. Let $\boldsymbol{T}^*$ be an optimal solution of the GW problem *i.e.* $GW_2(\mathbf{C}^1, \mathbf{C}^2, \boldsymbol{h}^1, \boldsymbol{h}^2) = \langle \mathbf{L}(\mathbf{C}^1, \mathbf{C}^2) \otimes \boldsymbol{T}^*, \boldsymbol{T}^* \rangle_F$. We define $\boldsymbol{M}(\boldsymbol{T}^*) := \mathbf{L}(\mathbf{C}^1, \mathbf{C}^2) \otimes \boldsymbol{T}^*$. We consider the problem:

$$\min_{\boldsymbol{T} \in \mathcal{U}(\boldsymbol{h}^1, \boldsymbol{h}^2)} \langle \boldsymbol{M}(\boldsymbol{T}^*), \boldsymbol{T} \rangle_F = \min_{\boldsymbol{T} \in \mathcal{U}(\boldsymbol{h}^1, \boldsymbol{h}^2)} \langle \mathbf{L}(\mathbf{C}^1, \mathbf{C}^2) \otimes \boldsymbol{T}^*, \boldsymbol{T} \rangle_F \tag{17}$$

We will first show that the optimal coupling for the Gromov-Wasserstein problem is also an optimal coupling for the problem equation 17, *i.e.* $\min_{\boldsymbol{T} \in \mathcal{U}(\boldsymbol{h}^1, \boldsymbol{h}^2)} \langle \boldsymbol{M}(\boldsymbol{T}^*), \boldsymbol{T} \rangle_F = \langle \boldsymbol{M}(\boldsymbol{T}^*), \boldsymbol{T}^* \rangle_F$. This result is based on the following theorem which relates a solution of a Quadratic Program (QP) with a solution of a Linear Program (LP):

Theorem 1 (Theorem 1.12 in (Murty, 1988)) *Consider the following (QP):*

$$\begin{array}{ll} \min_{\mathbf{x}} f(\mathbf{x}) & = \mathbf{c}\mathbf{x} + \mathbf{x}^T \mathbf{Q} \mathbf{x} \\ \textit{s.t.} & \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq 0 \end{array} \tag{18}$$

Then if $\mathbf{x}_*$ *is an optimal solution of equation 18 it is an optimal solution of the following (LP):*

$$\begin{array}{ll} \min_{\mathbf{x}} f(\mathbf{x}) & = (\mathbf{c} + \mathbf{x}_*^T \mathbf{Q}) \mathbf{x} \\ \textit{s.t.} & \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq 0 \end{array} \tag{19}$$

Applying Theorem 1 to our case gives exactly that:

$$\boldsymbol{T}^* \in \underset{\boldsymbol{T} \in \mathcal{U}(\boldsymbol{h}^1, \boldsymbol{h}^2)}{\arg\min} \langle \boldsymbol{M}(\boldsymbol{T}^*), \boldsymbol{T} \rangle_F \tag{20}$$

since $\boldsymbol{T}^*$ is an optimal solution of the GW problem and so $\min_{\boldsymbol{T} \in \mathcal{U}(\boldsymbol{h}^1, \boldsymbol{h}^2)} \langle \boldsymbol{M}(\boldsymbol{T}^*), \boldsymbol{T} \rangle_F = \langle \boldsymbol{M}(\boldsymbol{T}^*), \boldsymbol{T}^* \rangle_F$.

Now let $\boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{\beta}^*(\boldsymbol{T}^*)$ be an optimal solution to the dual problem of equation 17. Then by strong duality it implies that:

$$\min_{\boldsymbol{T} \in \mathcal{U}(\boldsymbol{h}^1, \boldsymbol{h}^2)} \langle \boldsymbol{M}(\boldsymbol{T}^*), \boldsymbol{T} \rangle_F = \langle \boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{h}^1 \rangle + \langle \boldsymbol{\beta}^*(\boldsymbol{T}^*), \boldsymbol{h}^2 \rangle = \langle \boldsymbol{M}(\boldsymbol{T}^*), \boldsymbol{T}^* \rangle_F \tag{21}$$

Since $\langle \boldsymbol{M}(\boldsymbol{T}^*), \boldsymbol{T}^* \rangle_F = GW_2(\mathbf{C}^1, \mathbf{C}^2, \boldsymbol{h}^1, \boldsymbol{h}^2)$ we have:

$$GW_2(\mathbf{C}^1, \mathbf{C}^2, \boldsymbol{h}^1, \boldsymbol{h}^2) = \langle \boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{h}^1 \rangle + \langle \boldsymbol{\beta}^*(\boldsymbol{T}^*), \boldsymbol{h}^2 \rangle \tag{22}$$

To prove Proposition 3 the objective is to show that $\boldsymbol{\beta}^*(\boldsymbol{T}^*)$ is a subgradient of $F : \mathbf{q} \to GW(\mathbf{C}^1, \mathbf{C}^2, \boldsymbol{h}^1, \mathbf{q})$ (by symmetry the result will be true for $\boldsymbol{\alpha}^*(\boldsymbol{T}^*)$). In other words we want to prove that:

$$\forall \mathbf{q} \in \Sigma_m, \langle \boldsymbol{\beta}^*(\boldsymbol{T}^*), \mathbf{q} \rangle - \langle \boldsymbol{\beta}^*(\boldsymbol{T}^*), \boldsymbol{h}^2 \rangle \leq F(\mathbf{q}) - F(\boldsymbol{h}^2) \tag{23}$$

This condition can be rewritten based on the following simple lemma:

Lemma 1 *The dual variable* $\boldsymbol{\beta}^*(\boldsymbol{T}^*)$ *is a subgradient of* $F : \mathbf{q} \to GW_2(\mathbf{C}^1, \mathbf{C}^2, \boldsymbol{h}^1, \mathbf{q})$ *if and only if:*

$$\forall \mathbf{q} \in \Sigma_m, \langle \boldsymbol{\beta}^*(\boldsymbol{T}^*), \mathbf{q} \rangle + \langle \boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{h}^1 \rangle \leq F(\mathbf{q}) \tag{24}$$

Proof. It is a subgradient if and only if:

$$\forall \mathbf{q} \in \Sigma_m, \langle \boldsymbol{\beta}^*(\boldsymbol{T}^*), \mathbf{q} \rangle - \langle \boldsymbol{\beta}^*(\boldsymbol{T}^*), \boldsymbol{h}^2 \rangle \leq F(\mathbf{q}) - F(\boldsymbol{h}^2) \tag{25}$$

However using equation 22 and the definition of F we have:

$$F(\boldsymbol{h}^2) = \langle \boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{h}^1 \rangle + \langle \boldsymbol{\beta}^*(\boldsymbol{T}^*), \boldsymbol{h}^2 \rangle \tag{26}$$

So overall:

$$\begin{aligned} & \langle \boldsymbol{\beta}^*(\boldsymbol{T}^*), \mathbf{q} \rangle - \langle \boldsymbol{\beta}^*(\boldsymbol{T}^*), \boldsymbol{h}^2 \rangle \leq F(\mathbf{q}) - (\langle \boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{h}^1 \rangle + \langle \boldsymbol{\beta}^*(\boldsymbol{T}^*), \boldsymbol{h}^2 \rangle) \\ \iff & \langle \boldsymbol{\beta}^*(\boldsymbol{T}^*), \mathbf{q} \rangle + \langle \boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{h}^1 \rangle \leq F(\mathbf{q}) \end{aligned} \tag{27}$$

□

In order to prove Proposition 3 we have to prove that the condition in Lemma 1 is satisfied. We will do so by leveraging the weak-duality of the GW problem as described in the next lemma:

Lemma 2 *For any vectors $\boldsymbol{\alpha} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^m$ we define:*

$$\mathcal{G}(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \min_{\boldsymbol{T} \geq 0} \langle \mathbf{L}(\mathbf{C}^1, \mathbf{C}^2) \otimes \boldsymbol{T} - \boldsymbol{\alpha} \mathbf{1}_m^\top - \mathbf{1}_n \boldsymbol{\beta}^\top, \boldsymbol{T} \rangle$$

Let $\boldsymbol{T}^*$ be an optimal solution of the GW problem. Consider:

$$\min_{\boldsymbol{T} \in \mathcal{U}(\boldsymbol{h}^1, \boldsymbol{h}^2)} \langle \boldsymbol{M}(\boldsymbol{T}^*), \boldsymbol{T} \rangle_F \tag{28}$$

where $\boldsymbol{M}(\boldsymbol{T}^*) := \mathbf{L}(\mathbf{C}^1, \mathbf{C}^2) \otimes \boldsymbol{T}^*$. Let $\boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{\beta}^*(\boldsymbol{T}^*)$ be the dual variables of the problem in equation 28. If $\mathcal{G}(\boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{\beta}^*(\boldsymbol{T}^*)) = 0$ then $\boldsymbol{\beta}^*(\boldsymbol{T}^*)$ is a subgradient of $F : \mathbf{q} \to GW_2(\mathbf{C}^1, \mathbf{C}^1, \boldsymbol{h}^1, \mathbf{q})$

Proof. Let $\mathbf{q} \in \Sigma_m$ be any weights vector be fixed. Recall that $F : \mathbf{q} \to GW_2(\mathbf{C}^1, \mathbf{C}^2, \boldsymbol{h}^1, \mathbf{q})$ so that:

$$F(\mathbf{q}) = GW_2(\mathbf{C}^1, \mathbf{C}^2, \boldsymbol{h}^1, \mathbf{q}) = \min_{\boldsymbol{T} \in \mathcal{U}(\boldsymbol{h}^1, \mathbf{q})} \langle \mathbf{L}(\mathbf{C}^1, \mathbf{C}^2) \otimes \boldsymbol{T}, \boldsymbol{T} \rangle \tag{29}$$

The Lagrangian associated to equation 29 reads:

$$\min_{\boldsymbol{T} \geq 0} \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \mathtt{L}(\boldsymbol{T}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \text{ where } \mathtt{L}(\boldsymbol{T}, \boldsymbol{\alpha}, \boldsymbol{\beta}) := \langle \mathbf{L}(\mathbf{C}^1, \mathbf{C}^2) \otimes \boldsymbol{T}, \boldsymbol{T} \rangle + \langle \boldsymbol{h}^1 - \boldsymbol{T} \mathbf{1}_m, \boldsymbol{\alpha} \rangle + \langle \mathbf{q} - \boldsymbol{T}^\top \mathbf{1}_n, \boldsymbol{\beta} \rangle \tag{30}$$

Moreover by weak Lagrangian duality:

$$\min_{\boldsymbol{T} \geq 0} \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \mathtt{L}(\boldsymbol{T}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \geq \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \min_{\boldsymbol{T} \geq 0} \mathtt{L}(\boldsymbol{T}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \tag{31}$$

However:

$$\begin{aligned} \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \min_{\boldsymbol{T} \geq 0} \mathtt{L}(\boldsymbol{T}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \langle \boldsymbol{\alpha}, \boldsymbol{h}^1 \rangle + \langle \boldsymbol{\beta}, \mathbf{q} \rangle + \min_{\boldsymbol{T} \geq 0} \langle \mathbf{L}(\mathbf{C}^1, \mathbf{C}^2) \otimes \boldsymbol{T} - \boldsymbol{\alpha} \mathbf{1}_m^\top - \mathbf{1}_n \boldsymbol{\beta}^\top, \boldsymbol{T} \rangle \\ &= \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \langle \boldsymbol{\alpha}, \boldsymbol{h}^1 \rangle + \langle \boldsymbol{\beta}, \mathbf{q} \rangle + \mathcal{G}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \end{aligned}$$

So by considering the dual variable $\boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{\beta}^*(\boldsymbol{T}^*)$ defined previously we have:

$$\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \min_{\boldsymbol{T} \geq 0} \mathtt{L}(\boldsymbol{T}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \geq \langle \boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{h}^1 \rangle + \langle \boldsymbol{\beta}^*(\boldsymbol{T}^*), \mathbf{q} \rangle + \mathcal{G}(\boldsymbol{\alpha}^*(\boldsymbol{\pi}^*), \boldsymbol{\beta}^*(\boldsymbol{T}^*)) \tag{32}$$

Now combining equation 31 and equation 32 we have:

$$\min_{\boldsymbol{T} \geq 0} \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \mathtt{L}(\boldsymbol{T}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \geq \langle \boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{h}^1 \rangle + \langle \boldsymbol{\beta}^*(\boldsymbol{T}^*), \mathbf{q} \rangle + \mathcal{G}(\boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{\beta}^*(\boldsymbol{T}^*)) \tag{33}$$

Since $F(\mathbf{q}) = \min_{\boldsymbol{T} \geq 0} \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \mathtt{L}(\boldsymbol{T}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ we have proven that:

$$\forall \mathbf{q} \in \Sigma_m, \langle \boldsymbol{\beta}^*(\boldsymbol{T}^*), \mathbf{q} \rangle + \langle \boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{h}^1 \rangle + \mathcal{G}(\boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{\beta}^*(\boldsymbol{T}^*)) \leq F(\mathbf{q}) \tag{34}$$

However Lemma 1 states that $\boldsymbol{\beta}^*(\boldsymbol{T}^*)$ is a subgradient of F if and only if:

$$\forall \mathbf{q} \in \Sigma_m, \langle \boldsymbol{\beta}^*(\boldsymbol{T}^*), \mathbf{q} \rangle + \langle \boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{h}^1 \rangle \leq F(\mathbf{q}) \tag{35}$$

So combining equation 34 with Lemma 1 proves:

$$\mathcal{G}(\boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{\beta}^*(\boldsymbol{T}^*)) \geq 0 \implies \boldsymbol{\beta}^*(\boldsymbol{T}^*) \text{ is a subgradient of } F \tag{36}$$

However we have $F(\boldsymbol{h}^2) = \langle \boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{h}^1 \rangle + \langle \boldsymbol{\beta}^*(\boldsymbol{T}^*), \boldsymbol{h}^2 \rangle$ by equation 26. So $\mathcal{G}(\boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{\beta}^*(\boldsymbol{T}^*)) \leq 0$ using equation 34 with $\mathbf{q} = \boldsymbol{h}^2$. So we can only hope to have $\mathcal{G}(\boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{\beta}^*(\boldsymbol{T}^*)) = 0$. □

The previous lemma states that it is sufficient to look at the quantity $\mathcal{G}(\boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{\beta}^*(\boldsymbol{T}^*))$ in order to prove that $\boldsymbol{\beta}^*(\boldsymbol{T}^*)$ is a subgradient of F. Interestingly the condition $\mathcal{G}(\boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{\beta}^*(\boldsymbol{T}^*)) = 0$ is satisfied which proves Proposition 3 as sated in the next lemma:

Lemma 3 *With previous notations we have $\mathcal{G}(\boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{\beta}^*(\boldsymbol{T}^*)) = 0$. In particular $\boldsymbol{\beta}^*(\boldsymbol{T}^*)$ is a subgradient of F so that Proposition 3 is valid.*

Proof. We want to find:

$$\mathcal{G}(\boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{\beta}^*(\boldsymbol{T}^*)) = \min_{\boldsymbol{T} \geq 0} \langle \mathbf{L}(\mathbf{C}^1, \mathbf{C}^2) \otimes \boldsymbol{T} - \boldsymbol{\alpha}^*(\boldsymbol{T}^*)\mathbf{1}_m^\top - \mathbf{1}_n \boldsymbol{\beta}^*(\boldsymbol{T}^*)^\top, \boldsymbol{T} \rangle$$

We define $H(\boldsymbol{T}) := \langle \mathbf{L}(\mathbf{C}_1, \mathbf{C}_2) \otimes \boldsymbol{T} - \boldsymbol{\alpha}^*(\boldsymbol{T}^*)\mathbf{1}_m^\top - \mathbf{1}_n \boldsymbol{\beta}^*(\boldsymbol{T}^*)^\top, \boldsymbol{T} \rangle$. Since $\boldsymbol{T}^*$ is optimal coupling for $\min_{\boldsymbol{T} \in \mathcal{U}(\boldsymbol{h}^1, \boldsymbol{h}^2)} \langle \boldsymbol{M}(\boldsymbol{T}^*), \boldsymbol{T} \rangle_F$ by equation 20 then for all i, j we have $T^*_{ij}(\boldsymbol{M}(\boldsymbol{T}^*)_{ij} - \alpha^*_i(\boldsymbol{T}^*) - \beta^*_j(\boldsymbol{T}^*)) = 0$ by the property of the optimal couplings for the Wasserstein problems. Equivalently:

$$\forall (i,j) \in [n] \times [m],\ T^*_{ij}([\mathbf{L}(\mathbf{C}^1, \mathbf{C}^2) \otimes \boldsymbol{T}^*]_{ij} - \alpha^*_i(\boldsymbol{T}^*) - \beta^*_j(\boldsymbol{T}^*)) = 0 \tag{37}$$

Then:

$$\begin{aligned} H(\boldsymbol{T}^*) &= \operatorname{tr}\left(\boldsymbol{T}^{*\top} (\mathbf{L}(\mathbf{C}^1, \mathbf{C}^2) \otimes \boldsymbol{T}^* - \boldsymbol{\alpha}^*(\boldsymbol{T}^*)\mathbf{1}_m^\top - \mathbf{1}_n \boldsymbol{\beta}^*(\boldsymbol{T}^*)^\top) \right) \\ &= \sum_{ij} T^*_{ij} (\mathbf{L}(\mathbf{C}^1, \mathbf{C}^2) \otimes \boldsymbol{T}^* - \boldsymbol{\alpha}^*(\boldsymbol{T}^*)\mathbf{1}_m^\top - \mathbf{1}_n \boldsymbol{\beta}^*(\boldsymbol{T}^*)^\top)_{ij} \\ &= \sum_{ij} T^*_{ij} ([\mathbf{L}(\mathbf{C}^1, \mathbf{C}^2) \otimes \boldsymbol{T}^*]_{ij} - \alpha^*_i(\boldsymbol{T}^*) - \beta^*_j(\boldsymbol{T}^*)) = 0 \end{aligned} \tag{38}$$

Which proves $\mathcal{G}(\boldsymbol{\alpha}^*(\boldsymbol{T}^*), \boldsymbol{\beta}^*(\boldsymbol{T}^*)) = 0$. □

3. Algorithmic details

3.1. GDL for graphs without attributes

We propose to model a graph as a weighted sum of pairwise relation matrices. More precisely, given a graph $G = (\boldsymbol{C}, \boldsymbol{h})$ and a *dictionary* $\{\overline{\boldsymbol{C}}_s\}_{s \in [S]} \subset S_N(\mathbb{R})$ we want to find a linear representation $\sum_{s \in [S]} w_s \overline{\boldsymbol{C}}_s$ of the graph G, as faithful as possible. The dictionary is made of pairwise relation matrices of graphs with order N. $\boldsymbol{w} = (w_s)_{s \in [S]} \in \Sigma_S$ is referred as *embedding* and denotes the coordinate of the graph G in the dictionary. We rely on the GW distance to assess the quality of our linear approximation and propose to minimize it to estimate its optimal embedding.

3.1.1. Gromov-Wasserstein unmixing

We first study the unmixing problem that consists in projecting a graph on the linear representation discussed above, *i.e.* estimate the optimal embedding $\boldsymbol{w}$ of a graph G. Our GW unmixing problem reads as

$$\min_{\boldsymbol{w} \in \Sigma_S} \quad GW_2^2\left(\boldsymbol{C}, \widetilde{\boldsymbol{C}}(\boldsymbol{w})\right) - \lambda \|\boldsymbol{w}\|_2^2 \tag{39}$$

$$\text{where,} \qquad \widetilde{\boldsymbol{C}}(\boldsymbol{w}) = \sum_s w_s \overline{\boldsymbol{C}_s} \tag{40}$$

where $\lambda \in \mathbb{R}^+$ induces a **negative** quadratic regularization promoting sparsity on the simplex as discussed in Li et al. (2016). In order to solve the non-convex problem in equation 39, we propose to use a Block Coordinate Descent (BCD) algorithms (Tseng, 2001). We fully detail the algorithm in the following and refer our readers to the main paper for the discussion on this approach.

Algorithm 1 BCD for GW unmixing problem 39

```
Initialize $\boldsymbol{w} = \frac{1}{S}\mathbf{1}_S$
repeat
    Compute OT matrix $\boldsymbol{T}$ of $GW_2^2\left(\boldsymbol{C}, \widetilde{\boldsymbol{C}}(\boldsymbol{w})\right)$, with CG algorithm (Vayer et al., 2018, Alg.1 & 2).
    Compute the optimal $\boldsymbol{w}$ solving equation 39 for a fixed $\boldsymbol{T}$ with CG algorithm 2
until convergence
```

Algorithm 2 CG for solving GW unmixing problem *w.r.t* $\boldsymbol{w}$ given $\boldsymbol{T}$

```
repeat
    Compute $\boldsymbol{g}$, gradients w.r.t $\boldsymbol{w}$ of $\mathcal{E}(\boldsymbol{C}, \widetilde{\boldsymbol{C}}(\boldsymbol{w}), \boldsymbol{T})$ following equation 42.
    Find direction $\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}\in\Sigma_S} \boldsymbol{x}^T\boldsymbol{g}$
    Line-search: denoting $\boldsymbol{z}(\gamma) = \gamma\boldsymbol{x}^\star + (1-\gamma)\boldsymbol{w}$,
        $$\gamma^\star = \underset{\gamma\in(0,1)}{\arg\min}\,\mathcal{E}(\boldsymbol{C}, \widetilde{\boldsymbol{C}}(\boldsymbol{z}(\gamma)), \boldsymbol{T}) = \underset{\gamma\in(0,1)}{\arg\min}\, a\gamma^2 + b\gamma + c \tag{41}$$
    $\boldsymbol{w} \leftarrow \boldsymbol{z}(\gamma^\star)$
until convergence
```

Partial derivates of the GW objective $\mathcal{E}$ *w.r.t* $\boldsymbol{w} = (\frac{\partial \mathcal{E}}{\partial w_s})_{s\in[S]}$ are expressed in equation 42, and further completed with gradient of the negative regularization term .

$$\frac{\partial \mathcal{E}}{\partial w_s}(\boldsymbol{C}, \widetilde{\boldsymbol{C}}(\boldsymbol{w}), \boldsymbol{T}) = 2tr\{\Big(\overline{\boldsymbol{C}_s} \odot \widetilde{\boldsymbol{C}}(\boldsymbol{w})\Big)\boldsymbol{h}\boldsymbol{h}^\top - \overline{\boldsymbol{C}_s}\boldsymbol{T}^\top\boldsymbol{C}^\top\boldsymbol{T}\} \tag{42}$$

The coefficient of the second-order polynom involved in equation 50 used to solve the problem, are expressed as follow,

$$a = tr\{\Big(\widetilde{\boldsymbol{C}}(\boldsymbol{x}^\star - \boldsymbol{w}) \odot \widetilde{\boldsymbol{C}}(\boldsymbol{x}^\star - \boldsymbol{w})\Big)\boldsymbol{h}\boldsymbol{h}^T\} - \lambda\|\boldsymbol{x}^\star - \boldsymbol{w}\|_2^2 \tag{43}$$

$$b = 2tr\{\Big(\widetilde{\boldsymbol{C}}(\boldsymbol{x}^\star - \boldsymbol{w}) \odot \widetilde{\boldsymbol{C}}(\boldsymbol{w})\Big)\boldsymbol{h}\boldsymbol{h}^\top - \widetilde{\boldsymbol{C}}(\boldsymbol{x}^\star - \boldsymbol{w})\boldsymbol{T}^\top\boldsymbol{C}^T\boldsymbol{T}\} - 2\lambda\langle\boldsymbol{w}, \boldsymbol{x} - \boldsymbol{w}\rangle \tag{44}$$

3.1.2. Dictionary Learning and online algorithm

Assume now that the dictionary $\{\overline{\boldsymbol{C}}_s\}_{s\in[S]}$ is not known and has to be estimated from the data. We define a dataset of K graphs $\big\{G^{(k)} : (\boldsymbol{C}^{(k)}, \boldsymbol{h}^{(k)})\big\}_{k\in[K]}$. Recall that each graph $G^{(k)}$ of order $N^{(k)}$ is summarized by its pairwise relation matrix $\boldsymbol{C}^{(k)} \in S_{N^{(k)}}(\mathbb{R})$ and weights $\boldsymbol{h}^{(k)} \in \Sigma_{N^{(k)}}$ over nodes. The DL problem, that aims at estimating the optimal dictionary for a given dataset can be expressed as:

$$\min_{\substack{\{\boldsymbol{w}^{(k)}\}_{k\in[K]} \\ \{\overline{\boldsymbol{C}}_s\}_{s\in[S]}}} \sum_{k=1}^{K} GW_2^2\left(\boldsymbol{C}^{(k)}, \widetilde{\boldsymbol{C}}(\boldsymbol{w}^{(k)})\right) - \lambda\|\boldsymbol{w}^{(k)}\|_2^2 \tag{45}$$

where $\boldsymbol{w}^{(k)} \in \Sigma_S, \overline{\boldsymbol{C}}_s \in S_N(\mathbb{R})$. We refer the reader to the main paper for the discussion on the non-convex problem 45. To tackle this problem we proposed a stochastic algorithm 3

Algorithm 3 GDL: stochastic update of atoms $\{\overline{\boldsymbol{C}}_s\}_{s\in[S]}$

```
Sample a minibatch of graphs $\mathcal{B} := \{\boldsymbol{C}^{(k)}\}_{k\in\mathcal{B}}$ .
Compute optimal $\{(\boldsymbol{w}^{(k)}, \boldsymbol{T}^{(k)})\}_{k\in[B]}$ by solving B independent unmixing problems with Alg.1.
Projected gradient step with estimated gradients $\widetilde{\nabla}_{\overline{\boldsymbol{C}}_s}$ (see equation 47), $\forall s \in [S]$:
    $$\overline{\boldsymbol{C}}_s \leftarrow Proj_{S_N(\mathbb{R})}(\overline{\boldsymbol{C}}_s - \eta_C\widetilde{\nabla}_{\overline{\boldsymbol{C}}_s}) \tag{46}$$
```

Estimated gradients *w.r.t* $\{\overline{\boldsymbol{C}_s}\}$ over a minibatch of graphs $\mathcal{B} := \{\boldsymbol{C}^{(k)}\}_{k\in\mathcal{B}}$ given unmixing solutions $\{(\boldsymbol{w}^{(k)}, \boldsymbol{T}^{(k)})\}_{k\in[B]}$ read:

$$\widetilde{\nabla}_{\overline{\boldsymbol{C}_s}}\left(\sum_{k\in\mathcal{B}}\mathcal{E}(\boldsymbol{C}^{(k)}, \widetilde{\boldsymbol{C}}(\boldsymbol{w}^{(k)}), \boldsymbol{T}^{(k)}\right) = \frac{2}{B}\sum_{k\in\mathcal{B}} w_s^{(k)}\{\widetilde{\boldsymbol{C}}(\boldsymbol{w}^{(k)}) \odot \boldsymbol{h}\boldsymbol{h}^\top - \boldsymbol{T}^{(k)\top}\boldsymbol{C}^{(k)\top}\boldsymbol{T}^{(k)}\} \tag{47}$$

3.2. GDL for graph with nodes attribute

We can also define the same DL procedure for labeled graphs using the FGW distance. The unmixing part defined in equation 39 can be adapted by considering a linear embedding of the similarity matrix *and* of the feature matrix parametrized by the *same* $\boldsymbol{w}$.

3.2.1. Fused Gromov-Wasserstein unmixing

More precisely, given a labeled graph $G = (\boldsymbol{C}, \boldsymbol{A}, \boldsymbol{h})$ (see Section 1) and a *dictionary* $\{(\overline{\boldsymbol{C}_s}, \overline{\boldsymbol{A}_s})\}_{s\in[S]} \subset S_N(\mathbb{R}) \times \mathbb{R}^{N\times d}$ we want to find a linear representation $(\sum_{s\in[S]} w_s\overline{\boldsymbol{C}_s}, \sum_{s\in[S]} w_s\overline{\boldsymbol{A}_s})$ of the labeled graph G, as faithful as possible in the sense of the FGW distance. The FGW unmixing problem that consists in projecting a labeled graph on the linear representation discussed above reads as follow, $\forall\alpha \in (0, 1)$,

$$\min_{\boldsymbol{w}\in\Sigma_S} \quad FGW_{2,\alpha}^2\left(\boldsymbol{C}, \boldsymbol{A}, \widetilde{\boldsymbol{C}}(\boldsymbol{w}), \widetilde{\boldsymbol{A}}(\boldsymbol{w})\right) - \lambda\|\boldsymbol{w}\|_2^2 \tag{48}$$

$$\text{where,} \quad \widetilde{\boldsymbol{C}}(\boldsymbol{w}) = \sum_s w_s\overline{\boldsymbol{C}_s} \quad \text{and} \quad \widetilde{\boldsymbol{A}}(\boldsymbol{w}) = \sum_s w_s\overline{\boldsymbol{A}_s} \tag{49}$$

where $\lambda \in \mathbb{R}^+$. A similar discussion than for the GW unmixing problem 39 holds. We adapt the BCD algorithm detailed in 1 to labeled graphs in Alg.4, to solve the non-convex problem of equation 48.

Algorithm 4 BCD for FGW unmixing problem 48

```
1: Initialize $\boldsymbol{w} = \frac{1}{S}\boldsymbol{1}_S$
2: repeat
3:     Compute OT matrix $\boldsymbol{T}$ of $FGW_{2,\alpha}^2\left(\boldsymbol{C}, \boldsymbol{A}, \widetilde{\boldsymbol{C}}(\boldsymbol{w}), \widetilde{\boldsymbol{A}}(\boldsymbol{w})\right)$, with CG algorithm (Vayer et al., 2018, Alg.1 & 2).
4:     Compute the optimal $\boldsymbol{w}$ solving equation 48 for a fixed $\boldsymbol{T}$ with CG algorithm 5.
5: until convergence
```

Algorithm 5 CG for solving FGW unmixing problem *w.r.t* $\boldsymbol{w}$ given $\boldsymbol{T}$

```
1: repeat
2:     Compute $\boldsymbol{g}$, gradients w.r.t $\boldsymbol{w}$ of equation 48 given $\boldsymbol{T}$ following equation 51.
3:     Find direction $\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}\in\Sigma_S} \boldsymbol{x}^T\boldsymbol{g}$
4:     Line-search: denoting $\boldsymbol{z}(\gamma) = \gamma\boldsymbol{x}^\star + (1-\gamma)\boldsymbol{w}$,
           $$\gamma^\star = \underset{\gamma\in(0,1)}{\arg\min}\, \alpha\mathcal{E}(\boldsymbol{C}, \widetilde{\boldsymbol{C}}(\boldsymbol{z}(\gamma)), \boldsymbol{T}) + (1-\alpha)\mathcal{F}(\boldsymbol{A}, \widetilde{\boldsymbol{A}}(\boldsymbol{z}(\gamma)), \boldsymbol{T}) = \underset{\gamma\in(0,1)}{\arg\min}\, a\gamma^2 + b\gamma + c \tag{50}$$
5:     $\boldsymbol{w} \leftarrow \boldsymbol{z}(\gamma^\star)$
6: until convergence
```

Partial derivates of the FGW objective $\mathcal{G}_\alpha := \alpha\mathcal{E} + (1-\alpha)\mathcal{F}$ *w.r.t* $\boldsymbol{w}$ are expressed in equations 42 and 51, and further completed with gradient of the negative regularization term.

$$\begin{aligned}\frac{\partial\mathcal{G}_\alpha}{\partial w_s}(\boldsymbol{C}, \boldsymbol{A}, \widetilde{\boldsymbol{C}}(\boldsymbol{w}), \widetilde{\boldsymbol{A}}(\boldsymbol{w}), \boldsymbol{T}) &= \alpha\frac{\partial\mathcal{E}}{\partial w_s}(\boldsymbol{C}, \widetilde{\boldsymbol{C}}(\boldsymbol{w}), \boldsymbol{T}) + (1-\alpha)\frac{\partial\mathcal{F}}{\partial w_s}(\boldsymbol{A}, \widetilde{\boldsymbol{A}}(\boldsymbol{w}), \boldsymbol{T}) \\ &= \alpha\frac{\partial\mathcal{E}}{\partial w_s}(\boldsymbol{C}, \widetilde{\boldsymbol{C}}(\boldsymbol{w}), \boldsymbol{T}) + 2(1-\alpha)tr\{\boldsymbol{D_h}\widetilde{\boldsymbol{A}}(\boldsymbol{w})\overline{\boldsymbol{A}_s}^\top - \boldsymbol{T}^\top\boldsymbol{A}\overline{\boldsymbol{A}_s}^\top\}\end{aligned} \tag{51}$$

where $\boldsymbol{D_h} = diag(\boldsymbol{h})$. The coefficients of the second-order polynom involved in equation 50 used to solve the problem, satisfy the following equations,

$$a = \alpha tr\{\left(\widetilde{\boldsymbol{C}}(\boldsymbol{x}^\star - \boldsymbol{w}) \odot \widetilde{\boldsymbol{C}}(\boldsymbol{x}^\star - \boldsymbol{w})\right)\boldsymbol{h}\boldsymbol{h}^T\} + (1-\alpha)tr\{\boldsymbol{D_h}\widetilde{\boldsymbol{A}}(\boldsymbol{x}^\star - \boldsymbol{w})\widetilde{\boldsymbol{A}}(\boldsymbol{x} - \boldsymbol{w})^\top\} - \lambda\|\boldsymbol{x}^\star - \boldsymbol{w}\|_2^2 \tag{52}$$

$$\begin{aligned}b = 2\alpha tr\{&\left(\widetilde{\boldsymbol{C}}(\boldsymbol{x}^\star - \boldsymbol{w}) \odot \widetilde{\boldsymbol{C}}(\boldsymbol{w})\right)\boldsymbol{h}\boldsymbol{h}^\top - \widetilde{\boldsymbol{C}}(\boldsymbol{x}^\star - \boldsymbol{w})\boldsymbol{T}^\top\boldsymbol{C}^T\boldsymbol{T}\} \\ &+ (1-\alpha)tr\{\boldsymbol{D_h}\widetilde{\boldsymbol{A}}(\boldsymbol{x}^\star - \boldsymbol{w})\widetilde{\boldsymbol{A}}(\boldsymbol{w})^\top - \boldsymbol{T}^\top\boldsymbol{A}\widetilde{\boldsymbol{A}}(\boldsymbol{x}^\star - \boldsymbol{w})^\top\} - 2\lambda\langle\boldsymbol{w}, \boldsymbol{x} - \boldsymbol{w}\rangle\end{aligned} \tag{53}$$

3.2.2. Dictionary Learning and online algorithm

Assume now that the dictionary $\{(\overline{\boldsymbol{C}_s}, \overline{\boldsymbol{A}_s})\}_{s\in[S]}$ is not known and has to be estimated from the data. We define a dataset of K labeled graphs $\left\{G^{(k)} : (\boldsymbol{C}^{(k)}, \boldsymbol{A}^{(k)}, \boldsymbol{h}^{(k)})\right\}_{k\in[K]}$. Recall that each labeled graph $G^{(k)}$ of order $N^{(k)}$ is summarized by its pairwise relation matrix $\boldsymbol{C}^{(k)} \in S_{N^{(k)}}(\mathbb{R})$, its matrix of node features $\boldsymbol{A}^{(k)} \in \mathbb{R}^{N^{(k)}\times d}$ and weights $\boldsymbol{h}^{(k)} \in \Sigma_{N^{(k)}}$ over nodes. The DL problem, that aims at estimating the optimal dictionary for a given dataset can be expressed as:

$$\min_{\substack{\{\boldsymbol{w}^{(k)}\}_{k\in[K]} \\ \{(\overline{\boldsymbol{C}_s}, \overline{\boldsymbol{A}_s})\}_{s\in[S]}}} \sum_{k=1}^{K} FGW_{2,\alpha}^2 \left(\boldsymbol{C}^{(k)}, \boldsymbol{A}^{(k)}, \widetilde{\boldsymbol{C}}(\boldsymbol{w}^{(k)}), \widetilde{\boldsymbol{A}}(\boldsymbol{w}^{(k)})\right) - \lambda \|\boldsymbol{w}^{(k)}\|_2^2 \tag{54}$$

where $\boldsymbol{w}^{(k)} \in \Sigma_S, \overline{\boldsymbol{C}_s} \in S_N(\mathbb{R}), \overline{\boldsymbol{A}_s} \in \mathbb{R}^{N\times d}$. We refer the reader to the main paper for the discussion on the non-convex problem 45 which can be transposed to problem 54. To tackle this problem we proposed a stochastic algorithm 6

Algorithm 6 GDL: stochastic update of atoms $\{(\overline{\boldsymbol{C}_s}, \overline{\boldsymbol{A}_s})\}_{s\in[S]}$

```
1: Sample a minibatch of graphs $\mathcal{B} := \{(\boldsymbol{C}^{(k)}, \boldsymbol{A}^{(k)})\}_{k\in\mathcal{B}}$ .
2: Compute optimal $\{(\boldsymbol{w}^{(k)}, \boldsymbol{T}^{(k)})\}_{k\in[B]}$ by solving B independent unmixing problems with Alg.4.
3: Gradients step with estimated gradients $\widetilde{\nabla}_{\overline{\boldsymbol{C}_s}}$ (see equation 47), and $\widetilde{\nabla}_{\overline{\boldsymbol{A}_s}}$ (see equation 56), $\forall s \in [S]$. :
     $\overline{\boldsymbol{C}_s} \leftarrow Proj_{S_N(\mathbb{R})}(\overline{\boldsymbol{C}_s} - \eta_C \widetilde{\nabla}_{\overline{\boldsymbol{C}_s}})$    and    $\overline{\boldsymbol{A}_s} \leftarrow \overline{\boldsymbol{A}_s} - \eta_A \widetilde{\nabla}_{\overline{\boldsymbol{A}_s}}$    (55)
```

Estimated gradients *w.r.t* $\{\overline{\boldsymbol{C}_s}\}$ and $\{\overline{\boldsymbol{A}_s}\}$ over a minibatch of graphs $\mathcal{B} := \{(\boldsymbol{C}^{(k)}, \boldsymbol{A}^{(k)})\}_{k\in\mathcal{B}}$ given unmixing solutions $\{(\boldsymbol{w}^{(k)}, \boldsymbol{T}^{(k)})\}_{k\in[B]}$ can be computed separately. The ones related to the GW objective are described in equation 47, while the ones related to the Wasserstein objective satisfy equation 56:

$$\widetilde{\nabla}_{\overline{\boldsymbol{A}_s}} \left(\sum_{k\in\mathcal{B}} \mathcal{F}(\boldsymbol{A}^{(k)}, \widetilde{\boldsymbol{A}}(\boldsymbol{w}^{(k)}), \boldsymbol{T}^{(k)}) \right) = \frac{2}{B} \sum_{k\in\mathcal{B}} w_s^{(k)} \{\boldsymbol{D_h} \widetilde{\boldsymbol{A}}(\boldsymbol{w}^{(k)}) - \boldsymbol{T}^\top \boldsymbol{A}^{(k)}\} \tag{56}$$

3.3. Learning the graph structure and nodes distribution

Here we extend our GDL model defined in equation 45 and propose to learn atoms of the form $\{\overline{\boldsymbol{C}_s}, \overline{\boldsymbol{h}_s}\}_{s\in[S]}$. In this setting we have two independent dictionaries modeling the relative importance of the nodes with $\overline{\boldsymbol{h}_s} \in \Sigma_N$, and their pairwise relations through $\overline{\boldsymbol{C}_s}$. This dictionary learning problem reads:

$$\min_{\substack{\{(\boldsymbol{w}^{(k)}, \boldsymbol{v}^{(k)})\}_{k\in[K]} \\ \{(\overline{\boldsymbol{C}_s}, \overline{\boldsymbol{h}_s})\}_{s\in[S]}}} \sum_{k=1}^{K} GW_2^2 \left(\boldsymbol{C}^{(k)}, \widetilde{\boldsymbol{C}}(\boldsymbol{w}^{(k)}), \boldsymbol{h}^{(k)}, \widetilde{\boldsymbol{h}}(\boldsymbol{v}^{(k)})\right) - \lambda \|\boldsymbol{w}^{(k)}\|_2^2 - \mu \|\boldsymbol{v}^{(k)}\|_2^2 \tag{57}$$

where $\boldsymbol{w}^{(k)}, \boldsymbol{v}^{(k)} \in \Sigma_S$ are the structure and distribution embeddings and the linear models are defined as:

$$\forall k,\ \widetilde{\boldsymbol{h}}(\boldsymbol{v}^{(k)}) = \sum_s v_s^{(k)} \overline{\boldsymbol{h}_s}, \quad \widetilde{\boldsymbol{C}}(\boldsymbol{w}^{(k)}) = \sum_s w_s^{(k)} \overline{\boldsymbol{C}_s} \tag{58}$$

Here we exploit fully the GW formalism by estimating simultaneously the graph distribution $\widetilde{\boldsymbol{h}}$ and its geometric structure $\widetilde{\boldsymbol{C}}$. Optimization problem 57 can be solved by an adaptation of stochastic Algorithm 3. Indeed, in the light of the proposition 3, we can derive the following equation 59 between the input graph $(\boldsymbol{C}^{(k)}, \boldsymbol{h}^{(k)})$ and its embedded representation $\widetilde{\boldsymbol{C}}(\boldsymbol{w}^{(k)}$ and $\widetilde{\boldsymbol{h}}(\boldsymbol{v}^{(k)})$, given an optimal coupling $\boldsymbol{T}^{(k)}$ satisfying Proposition 3,

$$2\langle \boldsymbol{L}(\boldsymbol{C}^{(k)}, \widetilde{\boldsymbol{C}}(\boldsymbol{w}^{(k)})) \otimes \boldsymbol{T}^{(k)}, \boldsymbol{T}^{(k)} \rangle = \langle \boldsymbol{u}^{(k)}, \boldsymbol{h}^{(k)} \rangle + \langle \widetilde{\boldsymbol{u}}^{(k)}, \widetilde{\boldsymbol{h}}(\boldsymbol{v}^{(k)}) \rangle \tag{59}$$

where $\boldsymbol{u}^{(k)}, \widetilde{\boldsymbol{u}}^{(k)}$ are dual potentials of the induced linear OT problem.

First, with this observation we estimate the structure/node weights unmixings $(\boldsymbol{w}^{(k)}, \boldsymbol{v}^{(k)})$ for the graph $G^{(k)}$. We proposed the BCD algorithm 7 derived from the initial BCD 1. Note that the dual variables of the induced linear OT problems are centered to ensure numerical stability.

Algorithm 7 BCD for extended GW unmixing problem inherent to equation 57

```
1: Initialize embeddings such as $\boldsymbol{w} = \boldsymbol{v} = \frac{1}{S}\mathbf{1}_S$
2: repeat
3:     Compute OT matrix $\boldsymbol{T}$ of $GW_2^2\left(\boldsymbol{C}, \widetilde{\boldsymbol{C}}(\boldsymbol{w}), \boldsymbol{h}, \widetilde{\boldsymbol{h}}(\boldsymbol{v})\right)$, with CG algorithm (Vayer et al., 2018, Alg.1 & 2). From the finale iteration of CG, get dual potentials $(\boldsymbol{u}, \widetilde{\boldsymbol{u}})$ of the corresponding linear OT problem (see Proposition 3).
4:     Compute the optimal $\boldsymbol{v}$ by minimizing equation 59 w.r.t $\boldsymbol{v}$ given $\widetilde{\boldsymbol{u}}$ with a CG algorithm.
5:     Compute the optimal $\boldsymbol{w}$ solving equation 39 given $\boldsymbol{T}$ and $\boldsymbol{v}$ with CG algorithm 2.
6: until convergence
```

Second, now that we benefit from an algorithm to project any graph $G^{(k)} = (\boldsymbol{C}^{(k)}, \boldsymbol{h}^{(k)})$ onto the linear representations described in 58, we extend the stochastic algorithm 3. to the problem 57. This extension is described in algorithm 8.

Algorithm 8 extended GDL: stochastic update of atoms $\{(\overline{\boldsymbol{C}_s}, \overline{\boldsymbol{h}_s})\}_{s\in[S]}$

```
1: Sample a minibatch of graphs $\mathcal{B} := \{(\boldsymbol{C}^{(k)}, \boldsymbol{h}^{(k)})\}_{k\in\mathcal{B}}$ .
2: Compute optimal embeddings $\{(\boldsymbol{w}^{(k)}, \boldsymbol{v}^{(k)})\}_{k\in[B]}$ coming jointly with the set of OT variables $(\boldsymbol{T}^{(k)}, \boldsymbol{u}^{(k)}, \widetilde{\boldsymbol{u}}^{(k)})$ by solving B independent unmixing problems with Alg.7.
3: Projected gradient step with estimated gradients $\widetilde{\nabla}_{\overline{\boldsymbol{C}_s}}$ (see equation 47) and $\widetilde{\nabla}_{\overline{\boldsymbol{h}_s}}$ (see equation 61), $\forall s \in [S]$:
   $$\overline{\boldsymbol{C}_s} \leftarrow Proj_{S_N(\mathbb{R})}(\overline{\boldsymbol{C}_s} - \eta_C \widetilde{\nabla}_{\overline{\boldsymbol{C}_s}}) \quad \text{and} \quad \overline{\boldsymbol{h}_s} \leftarrow Proj_{\Sigma_N}(\overline{\boldsymbol{h}_s} - \eta_h \widetilde{\nabla}_{\overline{\boldsymbol{h}_s}}) \tag{60}$$
```

For a minibatch a graphs $\{\boldsymbol{C}_k, \boldsymbol{h}_k\}_{k\in[B]}$, once each unmixing problems are solved independently estimating unmixings $\{(\boldsymbol{w}^{(k)}, \boldsymbol{w}^{(k)})\}_k$ and the underlying OT matrix $\boldsymbol{T}^{(k)}$ associated with potential $\widetilde{\boldsymbol{u}}^{(k)}$, we perform simultaneously a projected gradient step update of $\{\overline{\boldsymbol{C}}_s\}_s$ and $\{\overline{\boldsymbol{h}}_s\}_s$. The estimated gradients of equation 57 *w.r.t* $\{\overline{\boldsymbol{h}_s}\}_s$ reads $\forall s \in [S]$,

$$\widetilde{\nabla}_{\overline{\boldsymbol{h}_s}\cdot} = \frac{1}{2B} \sum_{k\in[B]} v_s^{(k)} \widetilde{u}^{(k)} \tag{61}$$

4. Numerical experiments

4.1. Datasets

Table 1. Datasets descriptions

datasets	features	#graphs	#classes	mean #nodes	min #nodes	max #nodes	median #nodes	mean connectivity rate
IMDB-B	None	1000	2	19.77	12	136	17	55.53
IMDB-M	None	1500	3	13.00	7	89	10	86.44
MUTAG	{0..2}	188	2	17.93	10	28	17.5	14.79
PTC-MR	{0, .., 17}	344	2	14.29	2	64	13	25.1
BZR	$\mathbb{R}^3$	405	2	35.75	13	57	35	6.70
COX2	$\mathbb{R}^3$	467	2	41.23	32	56	41	5.24
PROTEIN	$\mathbb{R}^{29}$	1113	2	29.06	4	620	26	23.58
ENZYMES	$\mathbb{R}^{18}$	600	6	32.63	2	126	32	17.14

We considered well-known benchmark datasets divided into three categories: i) IMDB-B and IMDB-M (Yanardag & Vishwanathan, 2015) gather graphs without node attributes derived from social networks; ii) graphs with discrete attributes representing chemical compounds from MUTAG (Debnath et al., 1991) and cuneiform signs from PTC-MR (Krichene et al., 2015); iii) graphs with real vectors as attributes, namely BZR, COX2 (Sutherland et al., 2003) and PROTEINS, ENZYMES (Borgwardt & Kriegel, 2005). Details on each dataset are reported in Table 1

4.2. Settings

In the following, we detail the benchmark of our methods on supervised classification along additional (shared) considerations we made regarding the learning of our models. To consistently benchmark methods and configurations, as real graph datasets commonly used in machine learning literature show a high variance considering structure, we perform a nested cross validation (using 9 folds for training, 1 for testing, and reporting the average accuracy of this experiment repeated 10 times) by keeping same folds across methods. All splits are balanced *w.r.t* labels. In following results, parameters of SVM are cross validated within $C \in \{10^{-7}, 10^{-6}, ..., 10^{7}\}$ and $\gamma \in \{2^{-10}, 2^{-9}, ..., 2^{10}\}$.

For our approach, similar dictionaries are considered for unsupervised classification presented in the main paper, than for the supervised classification benchmark detailed in the following. So we refer the reader to the main paper for most implementation details. For completeness, we picked a batch size of 16. We initialized learning rate on the structure $\{\overline{C}_s\}$ at 0.1. In the presence of node features, we set a learning rate on $\{\overline{A}_s\}$ of 0.1 if $\alpha < 0.5$ and 1.0 otherwise. We optimized our dictionaries without features over 20 epochs and those with features over 40 epochs. In the following, we denote GDL-w the SVMs derived from embeddings $\boldsymbol{w}$ endowed with the Mahalanobis distance. While GDL-g denotes the SVMs derived from embedded graphs with the (F)GW distance. (Xu, 2020) proposed a supervised extension to their Gromov-Wasserstein Factorization (GWF), we refer to GWF-r and GWF-f when the dictionary atoms have random size or when we fix it to match our method. His supervised approach consists in balancing the dictionary objective with a classification loss by plugging a MLP classifier to the unconstrained embedding space. We explicitly regularized the learning procedure by monitoring the accuracy on train splits. Note that in their approach they relaxed constraints of their unmixing problems by applying a softmax on unconstrained embeddings to conduct barycenters estimation. Moreover, they constrain the graph atoms to be non-negative as it enhances numerical stability of their learning procedure. For fair comparisons, we considered this restriction for all dictionaries even if we did not observe any noticeable impact of this hypothesis on our approach. As for unsupervised experiments, we followed their architecture choices. We further validated their regularization coefficient in $\{1., 0.1, 0.01, 0.001\}$. Their model converge over 10 epochs for datasets without features, and 20 epochs otherwise.

We also considered several kernel based approaches. (FGWK) The kernels $e^{-\gamma FGW}$ proposed by (Vayer et al., 2018) where pairwise distances are computed using CG algorithms using POT library (Flamary & Courty, 2017). To get a grasp of the approximation error from this algorithmic approach, we also applied the MCMC algorithm proposed by (Chowdhury & Needham, 2020) to compute FGW distance matrices with a better precision (S-GWK). As the proper graph representations for OT-based methods is still a question of key interest, we consistently benchmarked our approach and these kernels when we consider adjacency and shortest-path representations. Moreover, we experimented on the heat kernels over normalized laplacian matrices suggested by (Chowdhury & Needham, 2020) on datasets without attributes, where we validated the diffusion parameter $t \in \{5, 10, 20\}$. We also reproduced the benchmark for classification on Graph Kernels done by (Vayer et al., 2018) by keeping their tested parameters for each method. (SPK) denotes the shortest path kernel (Borgwardt & Kriegel, 2005), (RWK) the random walk kernel (Gärtner et al., 2003), (WLK) the Weisfeler Lehman kernel (Vishwanathan et al., 2010), (GK) the graphlet count kernel (Shervashidze et al., 2009). For real valued vector attributes, we consider the HOPPER kernel (HOPPERK) (Feragen et al., 2013) and the propagation kernel (PROPAK) (Neumann et al., 2016) . We built upon the GraKel library (Siglidis et al., 2020) to construct the kernels.

Finally to compare our performances to recent state-of-the-art models for supervised graph classification, we partly replicated the benchmark done by (Xu et al., 2018). We experimented on their best model GIN-0 and the model of (Niepert et al., 2016) PSCN. r. For both we used the Adam optimizer (Kingma & Ba, 2014) with initial learning rate 0.01 and decayed the learning rate by 0.5 every 50 epochs. The number of hidden units is chosen depending on dataset statistics as they propose, batch normalization (Ioffe & Szegedy, 2015) was applied on each of them. The batch size was fixed at 128. We fixed a dropout ratio of 0.5 after the dense layer (Srivastava et al., 2014). The number of epochs was 150 and the model with the best cross-validation accuracy averaged over the 10 folds was selected at each epoch.

4.3. Results on supervised classification

The accuracies of the nested-cross validation on described datasets are reported in Tables 2, 3, 4. First, we observe as anticipated that the model GIN-0 (Xu et al., 2018) outperforms most of the time other methods including PSCN, which has been consistently argued in their paper. Moreover, (F)GW kernels over the embedded graphs built thanks to our dictionary approach consistently outperforms (F)GW kernels from input graphs. Hence, it supports that our dictionaries are able to properly denoise and capture discriminant patterns of these graphs, outperforming other models expect GNN on 6 datasets out of 8. The Mahalanobis distance over embeddings $\boldsymbol{w}$ demonstrates satisfying results compared to FGWK relatively to the model simplification it brings. We also observe consistent improvements of the classification performances when we use the MCMC algorithm (Chowdhury & Needham, 2020) to estimate (F)GW pairwise distance matrices, for all tested graph representations reported. This estimation procedure for (F)GW distances is computationally heavy compared to the usual CG gradient algorithm (Vayer et al., 2018). Hence, we believe that it could bring significant improvements to our dictionary learning models but would increase too consequently the run time of solving unmixing problems required for each dictionary updates. Finally, results over adjacency and shortest path representations interestingly suggest that their suitability *w.r.t* (F)GW distance is correlated to the averaged connectivity rate (see 1) in different ways depending on the kind of node features. We envision to study these correlations in future works.

Table 2. **Graphs without attributes:** Classification results of 10-fold nested-cross validation on real datasets. Best results are highlighted in bolt independently of the depicted model category, and the best performances from not end-to-end supervised methods are reported in italic.

category	model	IMDB-B	IMDB-M
OT (Ours)	GDL-w (ADJ)	70.11(3.13)	49.01(3.66)
	GDL-g (ADJ)	*72.06(4.09)*	*50.64(4.41)*
	GDL-w (SP)	65.4(3.65)	48.03(3.80)
	GDL-g (SP)	68.24(4.38)	48.47(4.21)
OT	FGWK (ADJ)	70.8(3.54)	48.89(3.93)
	FGWK (SP)	65.0(3.69)	47.8(3.84)
	FGWK (heatLAP)	67.7(2.76)	48.11(3.96)
	S-GWK (ADJ)	71.95(3.87)	49.97(3.95)
	S-GWK (heatLAP)	71.05(3.02)	49.24(3.49)
	GWF-r (ADJ)	65.08(2.85)	47.53(3.16)
	GWF-f (ADJ)	64.68(2.27)	47.19(2.96)
Kernels	GK (K=3)	57.11(3.49)	41.85(4.52)
	SPK	56.18(2.87)	39.07(4.89)
GNN	PSCN	71.23(2.13)	45.7(2.71)
	GIN-0	**74.7(4.98)**	**52.19(2.71)**

Table 3. **Graphs with discrete attributes :** Classification results of 10-fold nested-cross validation on real datasets with discrete attributes (one-hot encoded). Best results are highlighted in bolt independently of the depicted model category, and the best performances from not end-to-end methods are reported in italic.

category	model	MUTAG	PTC-MR
OT (Ours)	GDL-w (ADJ)	81.07(7.81)	55.26(8.01)
	GDL-g (ADJ)	85.84(6.86)	58.45(7.73)
	GDL-w (SP)	84.58(6.70)	55.13(6.03)
	GDL-g (SP)	*87.09(6.34)*	57.09(6.59)
OT	FGWK (ADJ)	82.63(7.16)	56.17(8.85)
	FGWK (SP)	84.42(7.29)	55.4(6.97)
	S-GWK (ADJ)	84.08(6.93)	57.89(7.54)
	GWF-r (ADJ)	-	-
	GWF-f (ADJ)	-	-
Kernels	GK (K=3)	82.86(7.93)	57.11(7.24)
	SPK	83.29(8.01)	60.55(6.43)
	RWK	79.53(7.85)	55.71(6.86)
	WLK	86.44(7.95)	*63.14(6.59)*
GNN	PSCN	**91.4(4.41)**	58.9(5.12)
	GIN-0	88.95(4.91)	**64.12(6.83)**

Table 4. **Graphs with vectorial attributes:** Classification results of 10-fold nested-cross validation on real datasets with vectorial features. Best results are highlighted in bolt independently of the depicted model category, and the best performances from not end-to-end supervised methods are reported in italic.

category	model	BZR	COX2	ENZYMES	PROTEIN
OT (ours)	GDL-w (ADJ)	87.32(3.58)	76.59(3.18)	70.68(3.36)	72.13(3.14)
	GDL-g (ADJ)	*87.81(4.31)*	78.11(5.13)	71.44(4.19)	74.59(4.95)
	GDL-w (SP)	83.96(5.51)	75.9(3.81)	69.95(5.01)	72.95(3.68)
	GDL-g (SP)	84.61(5.89)	76.86(4.91)	71.47(5.98)	*74.86(4.38)*
OT	FGWK (ADJ)	85.61(5.17)	77.02(4.16)	72.17(3.95)	72.41(4.70)
	FGWK (SP)	84.15(6.39)	76.53(4.68)	70.53(6.21)	74.34(3.27)
	S-GWK (ADJ)	86.91(5.49)	77.85(4.35)	**73.03(3.84)**	73.51(4.96)
	GWF-r (ADJ)	83.61(4.96)	75.33(4.18)	72.53(5.39)	73.64(2.48)
	GWF-f (ADJ)	83.72(5.11)	74.96(4.0)	72.14(4.97)	73.06(2.06)
Kernels	HOPPERK	84.51(5.22)	*79.68(3.48)*	46.2(3.75)	72.07(3.06)
	PROPAK	80.01(5.11)	77.81(3.84)	71.84(5.80)	61.73(4.5)
GNN	PSCN	83.91(5.71)	75.21(3.29)	43.89(3.91)	74.96(2.71)
	GIN-0	**88.71(5.48)**	**81.13(4.51)**	68.6(3.69)	**76.31(2.94)**

4.4. Complementary results on unsupervised classification

vanilla GDL As mentioned in section 4 of the main paper, we considered a fixed batch size for learning our models on labeled graphs, which turned out to be a limitation for the dataset ENZYMES. We report in table 5 our models performance on this dataset for a batch size fixed to 64 instead of 32 within the framework detailed above. These results are consistent with those observed on the other datasets.

Table 5. Clustering : dataset ENZYMES

MODELS	ENZYMES
GDL	71.83(0.18)
GDL$_\lambda$	$\mathbf{72.92(0.28})$

extended version of GDL We report here a companion study for clustering tasks which further supports our extension of GDL to the learning of node weights. As there is no Mahalanobis upper-bound for the linear models learned with this extension as their node weights are a priori different, we compare performances of K-means with GW distance applied on the embedded graphs produced with vanilla GDL, the extended version of GDL denoted here GDL$_h$ and GWF. Similar considerations have been made for learning GDL$_{\boldsymbol{h}}$ than those detailed for GDL, and we completed these results with an ablation of the quadratic negative regularization parameterized by λ. Results provided in 6 show that GW Kmeans applied to the graph representations from our method GDL$_h$ leads to state-of-the-art performances.

Table 6. Clustering: RI from GW Kmeans on embedded graphs.

models	λ	IMDB-B	IMDB-M
GDL (ours)	0	51.54(0.29)	55.86(0.25)
	> 0	51.97(0.48)	56.41(0.35)
GDL$_{\mathbf{h}}$ (ours)	0	52.51(0.22)	$\mathbf{57.12(0.3})$
	> 0	$\mathbf{53.09(0.38})$	56.95(0.25)
GWF-r	NA	51.39(0.15)	55.80(0.21)
GWF-f	NA	50.93(0.39)	54.48(0.26)

4.5. Runtimes

We report in Table 7 averaged runtimes for the same relative precision of 10^{-4} to compute one graph embedding on learned dictionaries from real datasets.

Table 7. Averaged runtimes.

dataset	# atoms	GDL	GWF
IMDB-B	12	52 ms	123 ms
	16	69 ms	186 ms
IMDB-M	12	44 ms	101 ms
	18	71 ms	168 ms

References

Borgwardt, K. M. and Kriegel, H.-P. Shortest-path kernels on graphs. In *Fifth IEEE international conference on data mining (ICDM'05)*, pp. 8–pp. IEEE, 2005.

Chowdhury, S. and Needham, T. Generalized Spectral Clustering via Gromov-Wasserstein Learning. *arXiv:2006.04163 [cs, math, stat]*, June 2020. arXiv: 2006.04163.

Debnath, A. K., Lopez de Compadre, R. L., Debnath, G., Shusterman, A. J., and Hansch, C. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2):786–797, 1991.

Feragen, A., Kasenburg, N., Petersen, J., de Bruijne, M., and Borgwardt, K. M. Scalable kernels for graphs with continuous attributes. In *NIPS*, pp. 216–224, 2013.

Flamary, R. and Courty, N. Pot python optimal transport library. *GitHub: https://github. com/rflamary/POT*, 2017.

Gärtner, T., Flach, P., and Wrobel, S. On graph kernels: Hardness results and efficient alternatives. In *Learning theory and kernel machines*, pp. 129–143. Springer, 2003.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Krichene, W., Krichene, S., and Bayen, A. Efficient bregman projections onto the simplex. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pp. 3291–3298. IEEE, 2015.

Li, P., Rangapuram, S. S., and Slawski, M. Methods for sparse and low-rank recovery under simplex constraints. *arXiv preprint arXiv:1605.00507*, 2016.

Murty, K. *Linear Complementarity, Linear and Nonlinear Programming*. Sigma series in applied mathematics. Heldermann, 1988. ISBN 978-3-88538-403-8.

Neumann, M., Garnett, R., Bauckhage, C., and Kersting, K. Propagation kernels: efficient graph kernels from propagated information. *Machine Learning*, 102(2):209–245, 2016.

Niepert, M., Ahmed, M., and Kutzkov, K. Learning convolutional neural networks for graphs. In *International conference on machine learning*, pp. 2014–2023, 2016.

Peyré, G., Cuturi, M., and Solomon, J. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pp. 2664–2672, 2016.

Shervashidze, N., Vishwanathan, S., Petri, T., Mehlhorn, K., and Borgwardt, K. Efficient graphlet kernels for large graph comparison. In *Artificial intelligence and statistics*, pp. 488–495. PMLR, 2009.

Siglidis, G., Nikolentzos, G., Limnios, S., Giatsidis, C., Skianis, K., and Vazirgiannis, M. Grakel: A graph kernel library in python. *Journal of Machine Learning Research*, 21(54):1–5, 2020.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Sutherland, J. J., O'brien, L. A., and Weaver, D. F. Spline-fitting with a genetic algorithm: A method for developing classification structure- activity relationships. *Journal of chemical information and computer sciences*, 43(6):1906–1915, 2003.

Tseng, P. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. Fused gromov-wasserstein distance for structured objects: theoretical foundations and mathematical properties. *arXiv preprint arXiv:1811.02834*, 2018.

Vayer, T., Courty, N., Tavenard, R., and Flamary, R. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pp. 6275–6284. PMLR, 2019.

Vishwanathan, S. V. N., Schraudolph, N. N., Kondor, R., and Borgwardt, K. M. Graph kernels. *The Journal of Machine Learning Research*, 11:1201–1242, 2010.

Xu, H. Gromov-wasserstein factorization models for graph clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6478–6485, 2020.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Yanardag, P. and Vishwanathan, S. Deep graph kernels. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1365–1374, 2015.