Distributional Reduction: Unifying Dimensionality Reduction and Clustering with Gromov-Wasserstein Projection

Hugues Van Assel [* 1]   **Cédric Vincent-Cuaz** [* 2]   **Nicolas Courty** [3]   **Rémi Flamary** [4]   **Pascal Frossard** [2]
Titouan Vayer [5]

Abstract

Unsupervised learning aims to capture the underlying structure of potentially large and high-dimensional datasets. Traditionally, this involves using dimensionality reduction methods to project data onto lower-dimensional spaces or organizing points into meaningful clusters. In practice, these methods are used sequentially, without guaranteeing that the clustering aligns well with the conducted dimensionality reduction. In this work, we offer a fresh perspective: that of distributions. Leveraging tools from optimal transport, particularly the Gromov-Wasserstein distance, we unify clustering and dimensionality reduction into a single framework called distributional reduction. This allows us to jointly address clustering and dimensionality reduction with a single optimization problem. Through comprehensive experiments, we highlight the versatility of our method and show that it outperforms existing approaches across a variety of image and genomics datasets.

1. Introduction

One major objective of unsupervised learning (Hastie et al., 2009) is to provide interpretable and meaningful approximate representations of the data that best preserve its structure *i.e.* the underlying geometric relationships between the data samples. Similar in essence to Occam's principle frequently employed in supervised learning, the preference for unsupervised data representation often aligns with the pursuit of simplicity, interpretability or visualizability in the associated model. These aspects are determinant in many real-world applications, such as cell biology (Cantini et al.,

---

[*]Equal contribution [1]ENS de Lyon, UMPA UMR 5669 [2]EPFL, Lausanne LTS4 [3]Université Bretagne Sud, IRISA UMR 6074 [4]École polytechnique, IP Paris, CMAP UMR 7641 [5]Univ. Lyon, ENS de Lyon, UCBL, Inria LIP UMR 5668. Correspondence to: Hugues Van Assel <hugues.van_assel@ens-lyon.fr>, Cédric Vincent-Cuaz <cedric.vincent-cuaz@epfl.ch>.
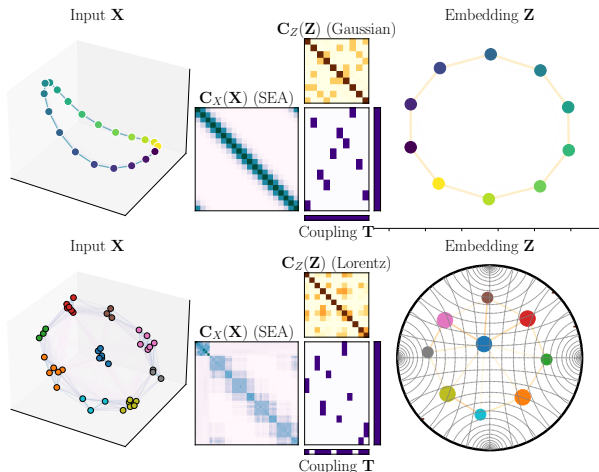
Figure 1: **Illustration of our DistR method** on two toys examples: points arranged on a circle (*top row*) and clusters with varying sizes (*bottom*) both in 3 dimensions. *Middle column*: input similarity matrix $\mathbf{C}_X(\mathbf{X})$ and final resulting embedding similarity $\mathbf{C}_Z(\mathbf{Z})$. Both are coupled through the coupling matrix $\mathbf{T}$ (depicted in purple, with its marginals)

2021; Ventre et al., 2023), where the interaction with domain experts is paramount for interpreting the results and extracting meaningful insights from the model.

Dimensionality reduction and clustering. When faced with the question of extracting interpretable representations, from a dataset $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)^\top \in \mathbb{R}^{N \times p}$ of N samples in $\mathbb{R}^p$, the machine learning community has proposed a variety of methods. Among them, dimensionality reduction (DR) algorithms have been widely used to summarize data in a low-dimensional space $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_N)^\top \in \mathbb{R}^{N \times d}$ with $d \ll p$, allowing for visualization of every individual points for small enough d (Agrawal et al., 2021; Van Der Maaten et al., 2009). Another major approach is to cluster the data into n groups, with n typically much smaller than N, and to summarize these groups through their centroids (Saxena et al., 2017; Ezugwu et al., 2022). Clustering is particularly interpretable since it provides a smaller number of points that can be easily inspected. The cluster assignments can also be analyzed. Both DR and clustering follow a similar

philosophy of summarization and reduction of the dataset using a smaller size representation.

Two sides of the same coin. As a matter of fact, methods from both families share many similitudes, including the construction of a similarity graph between input samples. In clustering, many popular approaches design a reduced or coarsened version of the initial similarity graph while preserving some of its spectral properties (Von Luxburg, 2007; Schaeffer, 2007). In DR, the goal is to solve the inverse problem of finding low-dimensional embeddings that generate a similarity graph close to the one computed from input data points (Ham et al., 2004; Hinton & Roweis, 2002). Our work builds on these converging viewpoints and addresses the following question: *can DR and clustering be expressed in a common and unified framework ?*

A distributional perspective. To answer this question, we propose to look at both problems from a distributional point of view, treating the data as an empirical probability distribution $\mu = \frac{1}{N} \sum_i \delta_{\mathbf{x}_i}$. This enables us to consider statistical measures of similarity such as Optimal Transport (OT), which is at the core of our work. On the one hand, OT and clustering are strongly related. The celebrated K-means approach can be seen as a particular case of minimal Wasserstein estimator where a distribution of n Diracs is optimized *w.r.t* their weights and positions (Canas & Rosasco, 2012). Other connections between spectral clustering and the OT-based Gromov-Wasserstein (GW) distance have been recently developed in Chowdhury & Needham (2021); Chen et al. (2023); Vincent-Cuaz et al. (2022a). On the other hand, the link between DR and OT has not been explored yet. DR methods, when modeling data as distributions, usually focus on joint distribution between samples within each space separately, see *e.g.* Van Assel et al. (2023) or Lu et al. (2019). Consequently, they do not consider couplings to transport samples across spaces of varying dimensions.

Our contributions. In this paper, we propose to bridge this gap by proposing a novel and general distributional framework that encompasses both DR and clustering as special cases. We notably cast those problems as finding a reduced distribution that minimizes the GW divergence from the original empirical data distribution. Our method proceeds by first constructing an input similarity matrix $\mathbf{C}_X(\mathbf{X})$ that is matched with the embedding similarity $\mathbf{C}_Z(\mathbf{Z})$ through an OT coupling matrix $\mathbf{T}$. The latter establishes correspondences between input and embedding samples. We illustrate this principle in Figure 1 where one can notice that $\mathbf{C}_Z(\mathbf{Z})$ preserves the topology of $\mathbf{C}_X(\mathbf{X})$ with a reduced number of nodes. The adaptivity of our model that can select an effective number of cluster $< n$, is visible in the bottom plot, where only the exact number of clusters in the original data (9 out of the 12 initially proposed) is automatically recovered. Our method can operate in any embedding space,

which is illustrated by projecting in either a 2D Euclidean plane or a Poincaré ball as embedding spaces.

We show that this framework is versatile and allows to recover many popular DR methods such as the kernel PCA and neighbor embedding algorithms, but also clustering algorithms such as K-means and spectral clustering. We first prove in Section 3 that DR can be formulated as a GW projection problem under some conditions on the loss and similarity functions. We then propose in Section 4 a novel formulation of data summarization as a minimal GW estimator that allows to select both the dimensionality of the embedding d (DR) but also the number of Diracs n (Clustering). Finally, we show in section 5 the practical interest of our approach, which regularly outperforms its competitors for various joint DR/Clustering tasks.

Notations. The ith entry of a vector $\mathbf{v}$ is denoted as either v_i or $[\mathbf{v}]_i$. Similarly, for a matrix $\mathbf{M}$, M_{ij} and $[\mathbf{M}]_{ij}$ both denote its entry (i, j). S_N is the set of permutations of $[\![N]\!]$. $P_N(\mathbb{R}^d)$ refers to the set of discrete probability measures composed of N points of $\mathbb{R}^d$. Σ_N stands for the probability simplex of size N that is $\Sigma_N := \{\mathbf{h} \in \mathbb{R}_+^N \text{ s.t. } \sum_i h_i = 1\}$. In all the paper, $\log(\mathbf{M}), \exp(\mathbf{M})$ are to be understood element-wise. For $\mathbf{x} \in \mathbb{R}^N$, $\mathrm{diag}(\mathbf{x})$ denotes the diagonal matrix whose elements are the x_i.

2. Background on Dimensionality Reduction and Optimal Transport

We start by reviewing the most popular DR approaches and we introduce the Gromov-Wasserstein problem.

2.1. Unified View of Dimensionality Reduction

Let $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)^\top \in \mathbb{R}^{N \times p}$ be an input dataset. Dimensionality reduction focuses on constructing a low-dimensional representation or *embedding* $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_N)^\top \in \mathbb{R}^{N \times d}$, where $d < p$. The latter should preserve a prescribed geometry for the dataset encoded via a symmetric pairwise similarity matrix $\mathbf{C}_\mathbf{X} \in \mathbb{R}_+^{N \times N}$ obtained from $\mathbf{X}$. To this end, most popular DR methods optimize $\mathbf{Z}$ such that a certain pairwise similarity matrix in the output space matches $\mathbf{C}_\mathbf{X}$ according to some criteria. We subsequently introduce the functions

$$\mathbf{C}_X : \mathbb{R}^{N \times p} \to \mathbb{R}^{N \times N}, \mathbf{C}_Z : \mathbb{R}^{N \times d} \to \mathbb{R}^{N \times N}, \quad (1)$$

which define pairwise similarity matrices in the input and output space from embeddings $\mathbf{X}$ and $\mathbf{Z}$. The DR problem can be formulated quite generally as the optimization problem

$$\min_{\mathbf{Z} \in \mathbb{R}^{N \times d}} \sum_{(i,j) \in [\![N]\!]^2} L\left([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{ij}\right). \quad \text{(DR)}$$

where $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a loss that quantifies how similar are two points in the input space $\mathbb{R}^p$ compared to two points in the output space $\mathbb{R}^d$. Various losses are used, such as the quadratic loss $L_2(x, y) \coloneqq (x - y)^2$ or the Kullback-Leibler divergence $L_{\mathrm{KL}}(x, y) \coloneqq x \log(x/y) - x + y$. Below, we recall several popular methods that can be placed within this framework.

Spectral methods. When $\mathbf{C}_X(\mathbf{X})$ is a positive semi-definite matrix, eq. (DR) recovers spectral methods by choosing the quadratic loss $L = L_2$ and $\mathbf{C}_Z(\mathbf{Z}) = (\langle \mathbf{z}_i, \mathbf{z}_j \rangle)_{(i,j) \in [\![N]\!]^2}$ the matrix of inner products in the embedding space. Indeed, in this case, the objective value of eq. (DR) reduces to

$$\sum_{(i,j) \in [\![N]\!]^2} L_2([\mathbf{C}_X(\mathbf{X})]_{ij}, \langle \mathbf{z}_i, \mathbf{z}_j \rangle) = \|\mathbf{C}_X(\mathbf{X}) - \mathbf{Z}\mathbf{Z}^\top\|_F^2$$

where $\| \cdot \|_F$ is the Frobenius norm. This problem is commonly known as kernel Principal Component Analysis (PCA) (Schölkopf et al., 1997) and an optimal solution is given by $\mathbf{Z}^\star = (\sqrt{\lambda_1}\mathbf{v}_1, ..., \sqrt{\lambda_d}\mathbf{v}_d)^\top$ where λ_i is the i-th largest eigenvalue of $\mathbf{C}_X(\mathbf{X})$ with corresponding eigenvector $\mathbf{v}_i$ (Eckart & Young, 1936). As shown by Ham et al. (2004); Ghojogh et al. (2021), numerous dimension reduction methods can be categorized in this manner. This includes PCA when $\mathbf{C}_X(\mathbf{X}) = \mathbf{X}\mathbf{X}^\top$ is the matrix of inner products in the input space; (classical) multidimensional scaling (Borg & Groenen, 2005), when $\mathbf{C}_X(\mathbf{X}) = -\frac{1}{2}\mathbf{H}\mathbf{D}_\mathbf{X}\mathbf{H}$ with $\mathbf{D}_\mathbf{X}$ the matrix of squared euclidean distance between the points in $\mathbb{R}^p$ and $\mathbf{H} = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top$ is the centering matrix; Isomap (Tenenbaum et al., 2000), with $\mathbf{C}_X(\mathbf{X}) = -\frac{1}{2}\mathbf{H}\mathbf{D}_\mathbf{X}^{(g)}\mathbf{H}$ with $\mathbf{D}_\mathbf{X}^{(g)}$ the geodesic distance matrix; Laplacian Eigenmap (Belkin & Niyogi, 2003), with $\mathbf{C}_X(\mathbf{X}) = \mathbf{L}_\mathbf{X}^\dagger$ the pseudo-inverse of the Laplacian associated to some adjacency matrix $\mathbf{W}_\mathbf{X}$; but also Locally Linear Embedding (Roweis & Saul, 2000), and Diffusion Map (Coifman & Lafon, 2006) (for all of these examples we refer to Ghojogh et al. 2021, Table 1).

Neighbor embedding methods. An alternative group of methods relies on neighbor embedding techniques which consists in minimizing in $\mathbf{Z}$ the quantity

$$\sum_{(i,j) \in [\![N]\!]^2} L_{\mathrm{KL}}([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{ij}). \quad \text{(NE)}$$

Within our framework, this corresponds to eq. (DR) with $L = L_{\mathrm{KL}}$. The objective function of popular methods such as stochastic neighbor embedding (SNE) (Hinton & Roweis, 2002) or t-SNE (Van der Maaten & Hinton, 2008) can be derived from eq. (NE) with a particular choice of $\mathbf{C}_X, \mathbf{C}_Z$. For instance SNE and t-SNE both consider in the input space a symmetrized version of the entropic affinity (Vladymyrov & Carreira-Perpinan, 2013; Van Assel et al., 2023). In the embedding space, $\mathbf{C}_Z(\mathbf{Z})$ is usually

constructed from a "kernel" matrix $\mathbf{K}_\mathbf{Z}$ which undergoes a scalar (Van der Maaten & Hinton, 2008), row-stochastic (Hinton & Roweis, 2002) or doubly stochastic (Lu et al., 2019; Van Assel et al., 2023) normalization. Gaussian kernel $[\mathbf{K}_\mathbf{Z}]_{ij} = \exp(-\|\mathbf{z}_i - \mathbf{z}_j\|_2^2)$, or heavy-tailed Student-t kernel $[\mathbf{K}_\mathbf{Z}]_{ij} = (1 + \|\mathbf{z}_i - \mathbf{z}_j\|_2^2)^{-1}$, are typical choices (Van der Maaten & Hinton, 2008). We also emphasize that one can retrieve the UMAP objective (McInnes et al., 2018) from eq. (DR) using the binary cross-entropy loss. As described in (Van Assel et al., 2022) from a probabilistic point of view, all these objectives can be derived from a common Markov random field model with various graph priors.

Remark 2.1. The usual formulations of neighbor embedding methods rely on the loss $L(x, y) = x \log(x/y)$. However, due to the normalization, the total mass $\sum_{ij}[\mathbf{C}_Z(\mathbf{Z})]_{ij}$ is constant (often equal to 1) in all of the cases mentioned above. Thus the minimization in $\mathbf{Z}$ with the L_{KL} formulation is equivalent.

Hyperbolic geometry. The presented DR methods can also be extended to incorporate non-Euclidean geometries. Hyperbolic spaces (Chami et al., 2021; Fan et al., 2022; Guo et al., 2022; Lin et al., 2023) are of particular interest as they can capture hierarchical structures more effectively than Euclidean spaces and mitigate the curse of dimensionality by producing representations with lower distortion rates. For instance, Guo et al. (2022) adapted t-SNE by using the Poincaré distance and by changing the Student's t-distribution with a more general hyperbolic Cauchy distribution. Notions of projection subspaces can also be adapted, *e.g.* Chami et al. (2021) use horospheres as one-dimensional subspaces.

2.2. Optimal Transport Across Spaces

Optimal Transport (OT) (Villani et al., 2009; Peyré et al., 2019) is a popular framework for comparing probability distributions and is at the core of our contributions. We review in this section the Gromov-Wasserstein formulation of OT aiming at comparing distributions "across spaces".

Gromov-Wasserstein (GW). The GW framework (Mémoli, 2011; Sturm, 2012) comprises a collection of OT methods designed to compare distributions by examining the pairwise relations *within each domain*. For two matrices $\mathbf{C} \in \mathbb{R}^{N \times N}$, $\overline{\mathbf{C}} \in \mathbb{R}^{n \times n}$, and weights $\mathbf{h} \in \Sigma_N, \overline{\mathbf{h}} \in \Sigma_n$, the GW discrepancy is defined as

$$\mathrm{GW}_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{h}, \overline{\mathbf{h}}) \coloneqq \min_{\mathbf{T} \in \mathcal{U}(\mathbf{h}, \overline{\mathbf{h}})} E_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{T}),$$

$$\text{where} \quad E_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{T}) \coloneqq \sum_{ijkl} L(C_{ij}, \overline{C}_{kl}) T_{ik} T_{jl}, \quad \text{(GW)}$$

and $\mathcal{U}(\mathbf{h}, \overline{\mathbf{h}}) = \{\mathbf{T} \in \mathbb{R}_+^{N \times n} : \mathbf{T}\mathbf{1}_n = \mathbf{h}, \mathbf{T}^\top \mathbf{1}_N = \overline{\mathbf{h}}\}$ is the set of couplings between $\mathbf{h}$ and $\overline{\mathbf{h}}$. In this formulation, both pairs $(\mathbf{C}, \mathbf{h})$ and $(\overline{\mathbf{C}}, \overline{\mathbf{h}})$ can be interpreted as graphs

with corresponding connectivity matrices $\mathbf{C}, \overline{\mathbf{C}}$, and where nodes are weighted by histograms $\mathbf{h}, \overline{\mathbf{h}}$. Equation (GW) is thus a *quadratic problem* (in $\mathbf{T}$) which consists in finding a soft-assignment matrix $\mathbf{T}$ that aligns the nodes of the two graphs in a way that preserves their pairwise connectivities.

From a distributional perspective, GW can also be viewed as a distance between distributions that do not belong to the same metric space. For two discrete probability distributions $\mu_X = \sum_{i=1}^{N} [\mathbf{h}_X]_i \delta_{\mathbf{x}_i} \in \mathcal{P}_N(\mathbb{R}^p), \mu_Z = \sum_{i=1}^{n} [\mathbf{h}_Z]_i \delta_{\mathbf{z}_i} \in \mathcal{P}_n(\mathbb{R}^d)$ and pairwise similarity matrices $\mathbf{C}_X(\mathbf{X})$ and $\mathbf{C}_Z(\mathbf{Z})$ associated with the supports $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^\top$ and $\mathbf{Z} = (\mathbf{z}_1, \cdots, \mathbf{z}_n)^\top$, the quantity $\mathrm{GW}_L(\mathbf{C}_X(\mathbf{X}), \mathbf{C}_Z(\mathbf{Z}), \mathbf{h}_X, \mathbf{h}_Z)$ is a measure of dissimilarity or discrepancy between μ_X, μ_Z. Specifically, when $L = L_2$, and $\mathbf{C}_X(\mathbf{X}), \mathbf{C}_Z(\mathbf{Z})$ are pairwise distance matrices, GW defines a proper distance between μ_X and μ_Z with respect to measure preserving isometries[1] (Mémoli, 2011).

Due to its versatile properties, notably in comparing distributions over different domains, the GW problem has found many applications in machine learning, *e.g.,* for 3D meshes alignment (Solomon et al., 2016; Ezuz et al., 2017), NLP (Alvarez-Melis & Jaakkola, 2018), (co-)clustering (Peyré et al., 2016; Redko et al., 2020), single-cell analysis (Demetci et al., 2020), neuroimaging (Thual et al., 2022), graph representation learning (Xu, 2020; Vincent-Cuaz et al., 2021; Liu et al., 2022b; Vincent-Cuaz et al., 2022b; Zeng et al., 2023) and partitioning (Xu et al., 2019; Chowdhury & Needham, 2021).

In this work, we leverage the GW discrepancy to extend classical DR approaches, framing them as the projection of a distribution onto a space of lower dimensionality.

3. Dimensionality Reduction as OT

In this section, we present the strong connections between the classical eq. (DR) and the GW problem.

Gromov-Monge interpretation of DR. As suggested by eq. (DR), dimension reduction seeks to find embeddings $\mathbf{Z}$ so that the similarity between the (i, j) samples of the input data is as close as possible to the similarity between the (i, j) samples of the embeddings. Under reasonable assumptions about $\mathbf{C}_Z$, this also amounts to identifying the embedding $\mathbf{Z}$ *and* the best permutation that realigns the two similarity matrices. Recall that the function $\mathbf{C}_Z$ is equivariant by permutation, if, for any $N \times N$ permutation matrix $\mathbf{P}$ and any $\mathbf{Z}$, $\mathbf{C}_Z(\mathbf{PZ}) = \mathbf{P}\mathbf{C}_Z(\mathbf{Z})\mathbf{P}^\top$ (Bronstein et al., 2021). This type of assumption is natural for $\mathbf{C}_Z$: if we rearrange the order of samples (*i.e.* , the rows of $\mathbf{Z}$), we expect the similarity matrix between the samples to undergo the same

rearrangement.

Lemma 3.1. *Let $\mathbf{C}_Z$ be a permutation equivariant function and L any loss. The minimum eq. (DR) is equal to*

$$\min_{\mathbf{Z} \in \mathbb{R}^{N \times d}} \min_{\sigma \in S_N} \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{\sigma(i)\sigma(j)}) . \quad (2)$$

Also, any sol. $\mathbf{Z}$ of eq. (DR) is such that $(\mathbf{Z}, \mathrm{id})$ is sol. of eq. (2) and conversely any $(\mathbf{Z}, \sigma)$ sol. of eq. (2) is such that $\mathbf{Z}$ is a sol of eq. (DR) up to σ.

See proof in Appendix A.1. The correspondence established between eq. (DR) and eq. (2) unveils a *quadratic problem* similar to GW. Specifically, eq. (2) relates to the Gromov-Monge problem[2] (Mémoli & Needham, 2018) which seeks to identify, by solving a quadratic assignment problem (Cela, 2013), the permutation σ that best aligns two similarity matrices. Lemma 3.1 therefore shows that the best permutation is the identity when we also optimize the embedding $\mathbf{Z}$. We can delve deeper into these comparisons and demonstrate that the general formulation of dimension reduction is also equivalent to minimizing the Gromov-Wasserstein objective, which serves as a relaxation of the Gromov-Monge problem (Mémoli & Needham, 2022).

DR as GW Minimization. We suppose that the distributions have the same number of points ($N = n$) and uniform weights ($\mathbf{h}_Z = \mathbf{h}_X = \frac{1}{N}\mathbf{1}_N$). We recall that a matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$ is conditionally positive definite (CPD), *resp.* conditionally negative definite (CND), if it is symmetric and $\forall \mathbf{x} \in \mathbb{R}^N, \mathbf{x}^\top \mathbf{1}_N = 0$ s.t. $\mathbf{x}^\top \mathbf{C} \mathbf{x} \geq 0$, *resp.* ≤ 0.

We thus have the following theorem that extends Lemma 3.1 to the GW problem (proof can be found in Appendix A.2).

Theorem 3.2. *The minimum eq. (DR) is equal to $\min_{\mathbf{Z}} \mathrm{GW}_L(\mathbf{C}_X(\mathbf{X}), \mathbf{C}_Z(\mathbf{Z}), \frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N)$ in the following settings:*

(i) (spectral methods) $\mathbf{C}_X(\mathbf{X})$ is any matrix, $L = L_2$ and $\mathbf{C}_Z(\mathbf{Z}) = \mathbf{Z}\mathbf{Z}^\top$.

(ii) (neighbor embedding methods) $\mathrm{Im}(\mathbf{C}_X) \subseteq \mathbb{R}_{>0}^{N \times N}$, $L = L_{\mathrm{KL}}$, the matrix $\mathbf{C}_X(\mathbf{X})$ is CPD and, for any $\mathbf{Z}$,

$$\mathbf{C}_Z(\mathbf{Z}) = \mathrm{diag}(\boldsymbol{\alpha}_{\mathbf{Z}}) \mathbf{K}_{\mathbf{Z}} \mathrm{diag}(\boldsymbol{\beta}_{\mathbf{Z}}), \quad (3)$$

where $\boldsymbol{\alpha}_{\mathbf{Z}}, \boldsymbol{\beta}_{\mathbf{Z}} \in \mathbb{R}_{>0}^N$ and $\mathbf{K}_{\mathbf{Z}} \in \mathbb{R}_{>0}^{N \times N}$ is such that $\log(\mathbf{K}_{\mathbf{Z}})$ is CPD.

Remarkably, this result shows that *all spectral DR methods* can be seen as OT problems in disguise, as they all equivalently minimize a GW problem. The second point

---

[1]With weaker assumptions on $\mathbf{C}_X, \mathbf{C}_Z$, GW defines a pseudo-metric *w.r.t.* a different notion of isomorphism (Chowdhury & Mémoli, 2019). See Appendix B for more details.

[2]Precisely $\min_{\sigma \in S_N} \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{\sigma(i)\sigma(j)})$ is the Gromov-Monge discrepancy between two discrete distributions, with the same number of atoms and uniform weights.

of the theorem also provides some insights into this equivalence in the case of neighbor embedding methods. For instance, the Gaussian kernel $\mathbf{K_Z}$, used extensively in DR, satisfies the hypothesis as $\log(\mathbf{K_Z}) = (-\|\mathbf{z}_i - \mathbf{z}_j\|_2^2)_{ij}$ is CPD (see *e.g.* Maron & Lipman 2018). The terms $\boldsymbol{\alpha_Z}, \boldsymbol{\beta_Z}$ also allow for considering all the usual normalizations of $\mathbf{K_Z}$: by a scalar so as to have $\sum_{ij}[\mathbf{C}_Z(\mathbf{Z})]_{ij} = 1$, but also any row-stochastic or doubly stochastic normalization (with the Sinkhorn-Knopp algorithm Sinkhorn & Knopp 1967).

Matrices satisfying $\log(\mathbf{K_Z})$ being CPD are well-studied in the literature and are known as infinitely divisible matrices (Bhatia, 2006). It is noteworthy that the t-Student kernel does not fall into this category. Moreover, in the aforementioned neighbor embedding methods, the matrix $\mathbf{C}_X(\mathbf{X})$ is generally not CPD. The intriguing question of generalizing this result with weaker assumptions on $\mathbf{C}_Z$ and $\mathbf{C}_X$ remains open for future research. Interestingly, we have observed in the numerical experiments performed in Section 5 that the symmetric entropic affinity of Van Assel et al. (2023) was systematically CPD.

Remark 3.3. In Corollary A.4 of the appendix we also provide other sufficient conditions for neighbor embedding methods with the cross-entropy loss $L(x, y) = x\log(x/y)$. They rely on specific structures for $\mathbf{C}_Z$ but do not impose any assumptions on $\mathbf{C}_X$. Additionally, in Appendix A.3, we provide a *necessary* condition based on a bilinear relaxation of the GW problem. Although its applicability is limited due to challenges in proving it in full generality, it requires minimal assumptions on $\mathbf{C}_X, \mathbf{C}_Z$ and L.

In essence, both Lemma 3.1 and Theorem 3.2 indicate that dimensionality reduction can be reframed from a distributional perspective, with the search for an empirical distribution that aligns with the data distribution in the sense of optimal transport, through the lens of GW. In other words, DR is informally solving $\min_{\mathbf{z}_1, \cdots, \mathbf{z}_N} \mathrm{GW}(\frac{1}{n}\sum_{i=1}^N \delta_{\mathbf{x}_i}, \frac{1}{n}\sum_{i=1}^N \delta_{\mathbf{z}_i})$.

4. Distributional Reduction

The previous interpretation is significant because it allows for two generalizations. Firstly, beyond solely determining the positions $\mathbf{z}_i$ of Diracs (as in classical DR) we can now optimize *the mass* of the distribution μ_Z. This is interpreted as finding the relative importance of each point in the embedding $\mathbf{Z}$. More importantly, due to the flexibility of GW, we can also seek a distribution in the embedding with a smaller number of points $n < N$. This will result in both reducing the dimension *and* clustering the points in the embedding space through the optimal coupling. Informally, our *Distributional Reduction* (DistR) framework aims at solving $\min_{\mu_Z \in \mathcal{P}_n(\mathbb{R}^d)} \mathrm{GW}(\frac{1}{n}\sum_{i=1}^N \delta_{\mathbf{x}_i}, \mu_Z)$.

4.1. Distributional Reduction Problem

Precisely, the optimization problem that we tackle in this paper can be formulated as follows

$$\min_{\substack{\mathbf{Z} \in \mathbb{R}^{n \times d} \\ \mathbf{h}_Z \in \Sigma_n}} \mathrm{GW}_L(\mathbf{C}_X(\mathbf{X}), \mathbf{C}_Z(\mathbf{Z}), \mathbf{h}_X, \mathbf{h}_Z) \qquad \text{(DistR)}$$

This problem comes down to learning the closest graph $(\mathbf{C}_Z(\mathbf{Z}), \mathbf{h}_Z)$ parametrized by $\mathbf{Z}$ from $(\mathbf{C}_X(\mathbf{X}), \mathbf{h}_X)$ in the GW sense. When $n < N$, the embeddings $(\mathbf{z}_1, \cdots, \mathbf{z}_n)$ then act as *low-dimensional prototypical examples* of input samples, whose learned relative importance $\mathbf{h}_Z$ accommodates clusters of varying proportions in the input data $\mathbf{X}$ (see Figure 1). We refer to them as *prototypes*. The weight vector $\mathbf{h}_X$ is typically assumed to be uniform, that is $\mathbf{h}_X = \frac{1}{N}\mathbf{1}_N$, in the absence of prior knowledge. As discussed in Section 3, traditional DR amounts to setting $n = N, \mathbf{h}_Z = \frac{1}{N}\mathbf{1}_N$.

Clustering with DistR. One notable aspect of our model is its capability to simultaneously perform dimensionality reduction and clustering. Indeed, the optimal coupling $\mathbf{T} \in [0, 1]^{N \times n}$ of problem eq. (DistR) is, by construction, a soft-assignment matrix from the input data to the embeddings. It allows each point $\mathbf{x}_i$ to be linked to one or more prototypes $\mathbf{z}_j$ (clusters). In Section 4.2 we explore conditions where these soft assignments transform into hard ones, such that each point is therefore linked to a unique prototype/cluster.

A semi-relaxed objective. For a given embedding $\mathbf{Z}$ and $L = L_2$, it is known that minimizing in $\mathbf{h}_Z$ the DistR objective is *equivalent* to a problem that is computationally simpler than the usual GW one, namely the semi-relaxed GW divergence srGW_L (Vincent-Cuaz et al., 2022a):

$$\min_{\mathbf{T} \in \mathcal{U}_n(\mathbf{h}_X)} E_L(\mathbf{C}_X(\mathbf{X}), \mathbf{C}_Z(\mathbf{Z}), \mathbf{T}), \qquad \text{(srGW)}$$

where $\mathcal{U}_n(\mathbf{h}_X) := \{\mathbf{T} \in \mathbb{R}_+^{N \times n} : \mathbf{T}\mathbf{1}_n = \mathbf{h}_X\}$. To efficiently address eq. (DistR), we first observe that this equivalence holds for any inner divergence L with a straightforward adaptation of the proof in Vincent-Cuaz et al. (2022a). Additionally, we prove that srGW_L remains a divergence as soon as L is itself a divergence. Consequently, srGW_L vanishes iff both measures are isomorphic in a weak sense (Chowdhury & Mémoli, 2019). We emphasize that taking a proper divergence L is important (and basic assumptions on $\mathbf{X}$), as it avoids some trivial solutions as detailed in Appendix B.

Interestingly, srGW projections, *i.e.* optimizing only the weights $\mathbf{h}_Z$ over simple fixed supports $\mathbf{Z}$, have already remarkable representational capability. We illustrate this in Figure 2, by considering projections of a real-world dataset over 2D grids of increasing resolutions. Setting $\mathbf{C}_X(\mathbf{X}) = \mathbf{X}\mathbf{X}^\top$ and $\mathbf{C}_Z(\mathbf{X}) = \mathbf{Z}\mathbf{Z}^\top$, we can see that those projections recover faithful coarsened representations
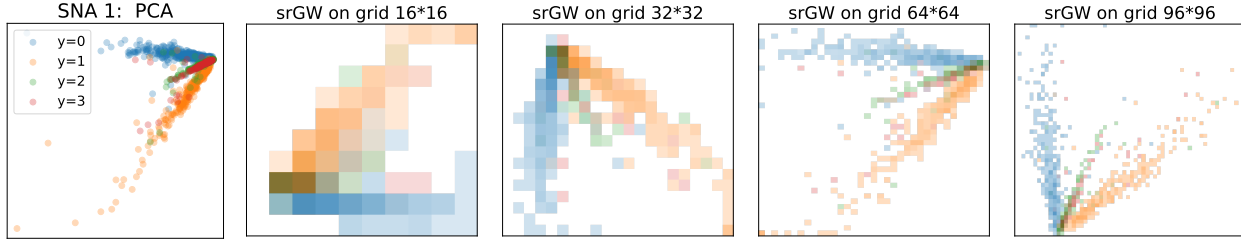
Figure 2: GW projections of a single-cell genomics dataset (Chen et al., 2019) on regular grids with increasing resolutions, respectively encoded as $C_X(\mathbf{X}) = \mathbf{X}\mathbf{X}^\top$ and $C_Z(\mathbf{Z}) = \mathbf{Z}\mathbf{Z}^\top$. Pixels on cropped grids are colored by interpolating samples' colors according to the transport plan and their intensity is proportional to their mass.

of the embeddings learned using PCA. DistR aims to exploit the full potential of this divergence by learning a few optimal prototypes that best represent the dataset.

4.2. Clustering Properties

In addition to the connections established between DistR and DR methods in Section 3, we elaborate now on the links between DistR and clustering methods.

In what follows, we call a coupling $\mathbf{T} \in [0,1]^{N \times n}$ with a single non-null element per row a *membership matrix*. When the coupling is a membership matrix each data point is associated with a single prototype thus achieving a hard clustering of the input samples.

We will see that a link can be drawn with kernel K-means using the analogy of *GW barycenters*. More precisely the *srGW barycenter* (Vincent-Cuaz et al., 2022a) seeks for a closest target graph $(\overline{\mathbf{C}}, \overline{\mathbf{h}})$ from $(\mathbf{C}_X, \mathbf{h}_X)$ by solving

$$\min_{\overline{\mathbf{C}} \in \mathbb{R}^{n \times n}} \min_{\mathbf{T} \in \mathcal{U}_n(\mathbf{h}_X)} E_L(\mathbf{C}_X(\mathbf{X}), \overline{\mathbf{C}}, \mathbf{T}) . \quad \text{(srGWB)}$$

We emphasize that the only (important) difference between eq. (srGWB) and eq. (DistR) is that there is no constraint imposed on $\overline{\mathbf{C}}$ in srGWB. In contrast, eq. (DistR) looks for minimizing over $\overline{\mathbf{C}} \in \{\mathbf{C}_Z(\mathbf{Z}) : \mathbf{Z} \in \mathbb{R}^{N \times d}\}$. For instance, choosing $\mathbf{C}_Z(\mathbf{Z}) = \mathbf{Z}\mathbf{Z}^\top$ in eq. (DistR) is equivalent to enforcing $\text{rank}(\overline{\mathbf{C}}) \leq d$ in eq. (srGWB).

We establish below that srGWB is of particular interest for clustering. The motivation for this arises from the findings of Chen et al. (2023), which demonstrate that when $\mathbf{C}_X(\mathbf{X})$ is positive semi-definite and $\mathbf{T}$ is *constrained* to belong to the set of membership matrices (as opposed to couplings in $\mathcal{U}_n(\mathbf{h})$), eq. (srGWB) is equivalent to a kernel K-means whose samples are weighted by $\mathbf{h}_X$ (Dhillon et al., 2004; 2007). These additional constraints are in fact unnecessary since we show below that the original srGWB problem admits membership matrices as the optimal coupling for a broader class of $\mathbf{C}_X(\mathbf{X})$ input matrices (see proofs in appendix C).

Theorem 4.1. *Let* $\mathbf{h}_X \in \Sigma_N$ *and* $L = L_2$. *Suppose that for any* $\mathbf{X} \in \mathbb{R}^{N \times p}$ *the matrix* $\mathbf{C}_X(\mathbf{X})$ *is CPD or CND.*

Then the problem eq. (srGWB) admits a membership matrix as optimal coupling, i.e. , there is a minimizer of $\mathbf{T} \in \mathcal{U}_n(\mathbf{h}_X) \to \min_{\overline{\mathbf{C}} \in \mathbb{R}^{n \times n}} E_L(\mathbf{C}_X(\mathbf{X}), \overline{\mathbf{C}}, \mathbf{T})$ *with only one non-zero value per row.*

Theorem 4.1 and relations proven in Chen et al. (2023) provide that eq. (srGWB) is equivalent to the aforementioned kernel K-means when $\mathbf{C}_X(\mathbf{X})$ is positive semi-definite. Moreover, as the (hard) clustering property holds for more generic types of matrices, namely CPD and CND, srGWB stands out as a fully-fledged clustering method. Although these results do not apply directly to DistR, we argue that they further legitimize the use of GW projections for clustering. Interestingly, we also observe in practice that the couplings obtained by DistR are always membership matrices, regardless of $\mathbf{C}_Z$. Further research will be carried out to better understand this phenomenon.

4.3. Computation

DistR is a non-convex problem that we propose to tackle using a Block Coordinate Descent algorithm (BCD, Tseng 2001) guaranteed to converge to local optimum (Grippo & Sciandrone, 2000; Lyu & Li, 2023). The BCD alternates between the two following steps. First, we optimize in $\mathbf{Z}$ for a fixed transport plan using gradient descent with adaptive learning rates (Kingma & Ba, 2014). Then we solve for a srGW problem given $\mathbf{Z}$. To this end, we benchmarked both the Conditional Gradient and Mirror Descent algorithms proposed in Vincent-Cuaz et al. (2022a), extended to support losses L_2 and L_{KL}.

Following Proposition 1 in (Peyré et al., 2016), a vanilla implementation leads to $\mathcal{O}(nN^2 + n^2N)$ operations to compute the loss or its gradient. In many DR methods, $\mathbf{C}_X(\mathbf{X})$ or $\mathbf{C}_Z(\mathbf{Z})$, or their transformations within the loss L, admit explicit low-rank factorizations. Including *e.g.* matrices involved in spectral methods and other similarity matrices derived from squared Euclidean distance matrices (Scetbon et al., 2022). In these settings, we exploit these factorizations to reduce the computational complexity of our solvers down to $\mathcal{O}(Nn(p+d) + (N+m)pd + n^2)$ when $L = L_2$,
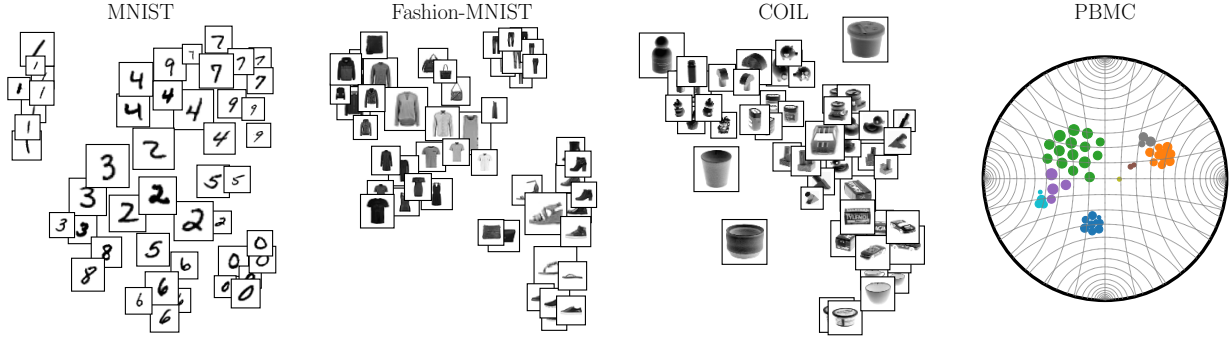
Figure 3: Examples of 2-dimensional embeddings produced by DistR using the SEA similarity for $\mathbf{C}_X$ and the Student's kernel for $\mathbf{C}_Z$. The latter computed in $\mathbb{R}^2$ for the first three datasets and the Poincaré ball for the last one. Displayed images are medoids for each cluster *i.e.* $\arg\max_i [\mathbf{C}_X(\mathbf{X})\mathbf{T}_{:,k}]_i$ for cluster k. The area of image k is proportional to $[\mathbf{h}_Z]_k$.

and $O(Nnd + n^2d)$ when $L = L_{\mathrm{KL}}$. We refer the reader interested in these algorithmic details to Appendix D.

4.4. Related Work

The closest to our work is the CO-Optimal-Transport (COOT) clustering approach proposed in (Redko et al., 2020) that estimates simultaneously a clustering of samples and features through the CO-Optimal Transport problem,

$$\min_{\substack{\mathbf{T}_1 \in \mathcal{U}(\mathbf{h}_1, \overline{\mathbf{h}}_1) \\ \mathbf{T}_2 \in \mathcal{U}(\mathbf{h}_2, \overline{\mathbf{h}}_2)}} \sum_{ijkl} (X_{ik} - Z_{jl})^2 [\mathbf{T}_1]_{ij}[\mathbf{T}_2]_{kl}, \quad \text{(COOT)}$$

where $\mathbf{h}_1 \in \Sigma_N$, $\overline{\mathbf{h}}_1 \in \Sigma_n$, $\mathbf{h}_2 \in \Sigma_p$ and $\overline{\mathbf{h}}_2 \in \Sigma_d$. We emphasize that COOT-clustering, which consists in optimizing the COOT objective above *w.r.t.* $\mathbf{Z}$, is a linear DR model as the reduction is done with the map $\mathbf{T}_2$. In contrast, DistR leverages the more expressive non-linear similarity functions of existing DR methods. Other joint DR-clustering approaches, such as (Liu et al., 2022a), involve modeling latent variables by a mixture of distributions. In comparison, our framework is more versatile, as it can easily adapt to any $(L, \mathbf{C}_X, \mathbf{C}_Z)$ of existing DR methods (Section 2.1).

5. Numerical Experiments

In this section, we demonstrate the relevance of our approach for joint clustering and dimensionality reduction.

Datasets. We experiment with popular labeled image datasets: COIL-20 (Nene et al., 1996), MNIST and fashion-MNIST (Xiao et al., 2017) as well as the following single-cell genomics datasets: PBMC 3k (Wolf et al., 2018), SNAREseq chromatin and gene expression (Chen et al., 2019) and the scRNA-seq dataset from Zeisel et al. (2015) with two hierarchical levels of label. Dataset sizes and pre-processing steps are detailed in appendix E.

Benchmarked methods. Let us recall that any DR method presented in Section 2.1 is fully characterized by a triplet

$(L, \mathbf{C}_X, \mathbf{C}_Z)$ of loss and pairwise similarity functions. Given $(L, \mathbf{C}_X, \mathbf{C}_Z)$, we compare our DistR model against sequential approaches of DR and clustering. Namely *DR then clustering* (DR→C) and *Clustering then DR* (C→DR) using the DR method associated with the same triplet. DR→C representations are constructed by first running the DR method (Section 2.1) associated with $(L, \mathbf{C}_X, \mathbf{C}_Z)$ thus obtaining an intermediate representation $\widetilde{\mathbf{Z}} \in \mathbb{R}^{N \times p}$. Then, spectral clustering (Von Luxburg, 2007) on the similarity matrix $\mathbf{C}_Z(\widetilde{\mathbf{Z}})$ is performed to compute a cluster assignment matrix $\widetilde{\mathbf{T}} \in \mathbb{R}^{N \times n}$. The final reduced representation in $\mathbb{R}^{n \times p}$ is the average of each point per cluster, *i.e.* the collection of the centroids, which is formally $\mathrm{diag}(\tilde{\mathbf{h}})^{-1}\widetilde{\mathbf{T}}^{\top}\widetilde{\mathbf{Z}} \in \mathbb{R}^{n \times p}$ where $\tilde{\mathbf{h}} = \widetilde{\mathbf{T}}^{\top}\mathbf{1}_N$. For C→DR, a cluster assignment matrix $\widehat{\mathbf{T}} \in \mathbb{R}^{N \times n}$ is first computed using spectral clustering on $\mathbf{C}_X(\mathbf{X})$. Then, the cluster centroid $\mathrm{diag}(\hat{\mathbf{h}})^{-1}\widehat{\mathbf{T}}^{\top}\mathbf{X}$, where $\hat{\mathbf{h}} = \widehat{\mathbf{T}}^{\top}\mathbf{1}_N$, is passed as input to the DR method associated with $(L, \mathbf{C}_X, \mathbf{C}_Z)$. Finally, we compare against the COOT-clustering model which enables to perform clustering and DR jointly but with many limitations compared to DistR as discussed in Section 4.4.

Implementation details.[1] Throughout, the spectral clustering implementation of `scikit-learn` (Pedregosa et al., 2011) is used to perform either the clustering steps or the initialization of transport plans. For all methods, $\mathbf{Z}$ is initialized from *i.i.d.* sampling of the standard gaussian distribution $\mathcal{N}(0, 1)$ (or an equivalent Wrapped Normal distribution in Hyperbolic spaces (Nagano et al., 2019)), and further optimized using `PyTorch`'s automatic differentiation (Paszke et al., 2017) with Adam optimizer (Kingma & Ba, 2014), or with RAdam in hyperbolic settings (Bécigneul & Ganea, 2018). OT-based solvers are built upon the `POT` (Flamary et al., 2021) library. and computations in Hyperbolic spaces are done with `Geoopt` (Kochurov et al., 2020). For the Hyperbolic space, all the computations were conducted in the Lorentz model (Nickel & Kiela, 2018), which is less

---

[1]Our code is provided with this submission.

	methods	$\mathbf{C}_X/\mathbf{C}_Z$	MNIST	FMNIST	COIL	SNA1	SNA2	ZEI1	ZEI2	PBMC
d = 10	DistR (ours)	$\langle,\rangle_{\mathbb{R}^p}$ / $\langle,\rangle_{\mathbb{R}^d}$	**76.9 (0.4)**	74.0 (0.6)	77.4 (0.7)	75.6 (0.0)	**86.6 (2.9)**	76.5 (1.0)	49.1 (1.9)	**86.3 (0.5)**
	DistR$_\varepsilon$ (ours)	-	76.5 (0.0)	**74.5 (0.0)**	78.5 (0.0)	68.9 (0.2)	78.9 (1.3)	79.8 (0.2)	**52.7 (0.2)**	85.1 (0.0)
	DR→C	-	73.9 (1.7)	63.9 (0.0)	70.6 (3.3)	**77.3 (2.5)**	66.9(14.3)	73.6 (0.8)	26.9 (7.9)	76.9 (1.2)
	C→DR	-	76.5 (0.0)	74.3 (0.0)	62.3 (0.0)	68.3 (0.2)	86.0 (0.3)	79.6 (0.1)	52.5 (0.1)	86.0 (0.0)
	COOT	NA	32.8 (2.5)	28.2 (6.0)	47.9 (1.0)	49.3 (6.8)	76.6 (6.5)	**81.0 (2.4)**	30.0 (1.6)	34.5 (1.3)
d = 2	DistR (ours)	SEA / St.	77.5 (0.8)	**76.8 (0.5)**	83.3 (0.7)	80.2 (1.9)	88.5 (0.0)	**81.0 (0.6)**	47.6 (1.1)	82.5 (1.2)
	DistR$_\varepsilon$ (ours)	-	**77.9 (0.2)**	75.6 (0.6)	**83.9 (0.5)**	**80.4 (1.8)**	**90.7 (0.1)**	79.9 (0.2)	**48.2 (0.6)**	86.4 (0.2)
	DR→C	-	74.8 (0.9)	75.7 (0.5)	80.3 (0.5)	77.2 (0.4)	89.3 (0.1)	79.0 (0.5)	47.4 (2.7)	82.0 (1.7)
	C→DR	-	76.2 (0.6)	75.0 (0.3)	81.6 (0.3)	77.6 (0.5)	89.8 (0.1)	78.8 (0.7)	45.8 (0.9)	**88.4 (0.5)**
	COOT	NA	26.1 (5.7)	24.9 (1.5)	42.5 (2.5)	32.6 (5.1)	56.8 (4.0)	78.2 (0.7)	25.2 (0.6)	28.9 (3.2)
d = 2	DistR (ours)	SEA / H-St.	**75.0 (0.0)**	**75.3 (0.6)**	70.2 (0.8)	75.4 (0.8)	**88.9 (1.8)**	77.4 (3.1)	**41.6 (1.7)**	**73.2 (1.6)**
	DistR$_\varepsilon$ (ours)	-	66.9 (0.5)	66.4 (0.3)	69.6 (0.9)	**81.3 (5.1)**	78.6 (1.0)	73.6 (1.8)	38.6 (0.7)	72.9 (2.3)
	DR→C	-	58.4 (4.6)	59.7 (9.7)	47.4 (1.5)	51.8 (5.9)	58.1 (9.9)	**80.6 (5.2)**	41.5 (2.4)	67.4 (7.2)
	C→DR	-	67.1 (1.8)	66.3 (0.5)	**71.1 (1.1)**	80.8 (3.9)	75.0(0.0)	71.2 (0.9)	37.5 (0.4)	67.6 (0.4)

Table 1: Average $\overline{\mathcal{SH}}$ scores (with standard deviation) for the main experiment (Section 5). The reported results are averaged on 5 different random seeds. The best result is in **bold**, second best is underlined. DistR refers to our solution relying on the conditional gradient to update $\mathbf{T}$ while DistR$_\varepsilon$ relies on KL mirror descent (details in Section 4.3 and Appendix D). Finally, St. refers to the Student's kernel and H-St. is its equivalent in the hyperbolic context.

prone to numerical errors. As such, we used the associated distance function to form $\mathbf{C}_Z$. After optimization, results are projected back to the Poincaré ball for visualization purpose. In this hyperbolic context, we adopted the formulation and the default hyperparameters of Guo et al. (2022) which generalize Student's t-distributios by Hyperbolic Cauchy distributions (denoted as H-St in the results Table 1).

Evaluation metric. We aim to evaluate both the DR and clustering abilities of our method DistR. For clustering evaluation, we use the `homogeneity_score` of `Torchmetrics` (Rosenberg & Hirschberg, 2007; Detlefsen et al., 2022) and denoted by $\mathcal{H}$. Relying on known class labels $\mathbf{y}$, this score ranging from 0 to 1 quantifies the extent to which each prototype $\mathbf{z}_k$ receives mass through $\mathbf{T}$ from points within the same class. To assess DR performance, we make use of the silhouette score (Rousseeuw, 1987) ranging from -1 to 1. This score is computed by first associating with each prototype $\mathbf{z}_k$ a label defined by the y maximizing $\sum_{i\in[\![N]\!]} T_{ik}\mathbf{1}_{y_i=y}$ where y_i is the class label of the input data point $\mathbf{x}_i$. Then it evaluates whether the relative position of prototypes, quantified via the appropriate metric depending on the embedding space, is aligned with the label they are assigned to. To compute the final silhouette, we slightly adapt the usual criterion to take into account the relative importance of each $\mathbf{z}_k$ given by the amount of mass it receives, that is $\sum_{i\in[\![N]\!]} T_{ik}$. More details can be found in Appendix E. We use the notation $\mathcal{S}$ for this score. To best quantify the trade-off between clustering and DR, our final general criterion is defined as the average of the homogeneity and silhouette score, where the latter is rescaled and

shifted such that it ranges from 0 to 1, *i.e.*

$$\overline{\mathcal{SH}} = \frac{1}{2}\left(\frac{\mathcal{S}+1}{2} + \mathcal{H}\right). \qquad \text{(Score)}$$

Setting and discussion. We experiment using both spectral and neighbor embedding methods (NE) for $(L, \mathbf{C}_X, \mathbf{C}_Z)$. For spectral, we consider the usual PCA setting which amounts to $\mathbf{C}_X(\mathbf{X}) = \mathbf{X}\mathbf{X}^\top$ and $\mathbf{C}_Z(\mathbf{Z}) = \mathbf{Z}\mathbf{Z}^\top$ choosing $d = 10$ as embedding dimension. Regarding NE, we rely on the *Symmetric Entropic Affinity* (SEA) from (Van Assel et al., 2023) for $\mathbf{C}_X$ and the scalar-normalized student similarity for $\mathbf{C}_Z$ (Van der Maaten & Hinton, 2008) in dimension $d = 2$. The SEA similarity enables controlling both mass and entropy making it robust to varying noise levels and depends on the perplexity hyperparameter ξ which determines the considered scale of dependency. For each configuration, we select the parameter ξ across the set $\{20, 50, 100, 150, 200, 250\}$, while the number of output samples n is validated in a set of 10 values, starting at the number of classes in the data and incrementing in steps of 20. For the computation of $\mathbf{T}$ in DistR (see Section 4.3), we benchmark our Conditional Gradient solver, and the Mirror Descent algorithm whose parameter ε is validated in the set $\{0.1, 1, 10\}$. $\overline{\mathcal{SH}}$ scores are displayed in Table 1. We also provide, in Appendix E, the corresponding separated scores $\mathcal{S}$ and $\mathcal{H}$. They show the superiority of DistR for 80% of the considered configurations. Therefore, our approach effectively achieves the ideal equilibrium between homogeneity and preservation of the structure of the prototypes. For a qualitative illustration, we display some embeddings produced by our method in Figure 3.

6. Conclusion

After making a connection between the GW problem and popular clustering and DR algorithms, we proposed a unifying framework denoted as DistR, enabling to jointly reduce the feature and sample sizes of an input dataset. We believe that the versatility of the GW framework will enable new extensions in unsupervised learning. For instance, the formalism associated with (semi-relaxed) GW barycenters naturally enables considering multiple inputs, potentially unaligned and of different sizes. Hence our approach can be useful to address challenges related to the multi-view dimensionality reduction and clustering problems. Other promising directions involve better capturing multiple dependency scales in the input data by hierarchically adapting the resolution of the embedding similarity graph.

Acknowledgments

References

Agrawal, A., Ali, A., Boyd, S., et al. Minimum-distortion embedding. *Foundations and Trends® in Machine Learning*, 14(3):211–378, 2021.

Alvarez-Melis, D. and Jaakkola, T. S. Gromov-wasserstein alignment of word embedding spaces. *arXiv preprint arXiv:1809.00013*, 2018.

Bécigneul, G. and Ganea, O.-E. Riemannian adaptive optimization methods. *arXiv preprint arXiv:1810.00760*, 2018.

Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Bhatia, R. Infinitely divisible matrices. *The American Mathematical Monthly*, 113(3):221–235, 2006.

Birkhoff, G. Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucuman, Ser. A*, 5:147–154, 1946.

Borg, I. and Groenen, P. J. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Canas, G. and Rosasco, L. Learning probability measures with respect to optimal transport metrics. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., and Baudot, A. Benchmarking joint multiomics dimensionality reduction approaches for the study of cancer. *Nature communications*, 12(1):124, 2021.

Cao, L., McLaren, D., and Plosker, S. Centrosymmetric stochastic matrices. *Linear and Multilinear Algebra*, 70 (3):449–464, 2022.

Cela, E. *The quadratic assignment problem: theory and algorithms*, volume 1. Springer Science & Business Media, 2013.

Chami, I., Gu, A., Nguyen, D., and Ré, C. Horopca: Hyperbolic dimensionality reduction via horospherical projections. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1419–1429. PMLR, 2021.

Chen, S., Lake, B. B., and Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, 37(12): 1452–1457, 2019.

Chen, Y., Yao, R., Yang, Y., and Chen, J. A gromov–wasserstein geometric view of spectrum-preserving graph coarsening. *arXiv preprint arXiv:2306.08854*, 2023.

Chowdhury, S. and Mémoli, F. The gromov–wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4): 757–787, 2019.

Chowdhury, S. and Needham, T. Generalized spectral clustering via gromov-wasserstein learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 712–720. PMLR, 2021.

Coifman, R. R. and Lafon, S. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

Demetci, P., Santorella, R., Sandstede, B., Noble, W. S., and Singh, R. Gromov-wasserstein optimal transport to align single-cell multi-omics data. *BioRxiv*, pp. 2020–04, 2020.

Detlefsen, N. S., Borovec, J., Schock, J., Jha, A. H., Koker, T., Di Liello, L., Stancl, D., Quan, C., Grechkin, M., and Falcon, W. Torchmetrics-measuring reproducibility in pytorch. *Journal of Open Source Software*, 7(70):4101, 2022.

Dhillon, I. S., Guan, Y., and Kulis, B. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 551–556, 2004.

Dhillon, I. S., Guan, Y., and Kulis, B. Weighted graph cuts without eigenvectors a multilevel approach. *IEEE transactions on pattern analysis and machine intelligence*, 29(11):1944–1957, 2007.

Eckart, C. and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., and Akinyelu, A. A. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110:104743, 2022.

Ezuz, D., Solomon, J., Kim, V. G., and Ben-Chen, M. Gwcnn: A metric alignment layer for deep shape analysis. In *Computer Graphics Forum*, volume 36, pp. 49–57. Wiley Online Library, 2017.

Fan, X., Yang, C.-H., and Vemuri, B. C. Nested hyperbolic spaces for dimensionality reduction and hyperbolic nn design. In *Conference on Computer Vision and Pattern Recognition*, pp. 356–365, June 2022.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., et al. Pot: Python optimal transport. *The Journal of Machine Learning Research*, 22(1):3571–3578, 2021.

Ghojogh, B., Ghodsi, A., Karray, F., and Crowley, M. Unified framework for spectral dimensionality reduction, maximum variance unfolding, and kernel learning by semidefinite programming: Tutorial and survey. *arXiv preprint arXiv:2106.15379*, 2021.

Grippo, L. and Sciandrone, M. On the convergence of the block nonlinear gauss–seidel method under convex constraints. *Operations research letters*, 26(3):127–136, 2000.

Guo, Y., Guo, H., and Yu, S. X. CO-SNE: Dimensionality reduction and visualization for hyperbolic data. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 11–20, Los Alamitos, CA, USA, jun 2022.

Ham, J., Lee, D. D., Mika, S., and Schölkopf, B. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 47, 2004.

Hastie, T., Tibshirani, R., and Friedman, J. *Unsupervised Learning*. Springer New York, New York, NY, 2009.

Hinton, G. E. and Roweis, S. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002.

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

Kochurov, M., Karimov, R., and Kozlukov, S. Geoopt: Riemannian optimization in pytorch, 2020.

Konno, H. A cutting plane algorithm for solving bilinear programs. *Mathematical Programming*, 11(1):14–27, 1976.

Lacoste-Julien, S. Convergence rate of frank-wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

Lin, Y.-W. E., Coifman, R. R., Mishne, G., and Talmon, R. Hyperbolic diffusion embedding and distance for hierarchical representation learning. In *International Conference on Machine Learning*, 2023.

Liu, W., Liao, X., Yang, Y., Lin, H., Yeong, J., Zhou, X., Shi, X., and Liu, J. Joint dimension reduction and clustering analysis of single-cell rna-seq and spatial transcriptomics data. *Nucleic acids research*, 50(12):e72–e72, 2022a.

Liu, W., Xie, J., Zhang, C., Yamada, M., Zheng, N., and Qian, H. Robust graph dictionary learning. In *The Eleventh International Conference on Learning Representations*, 2022b.

Lu, Y., Corander, J., and Yang, Z. Doubly stochastic neighbor embedding on spheres. *Pattern Recognition Letters*, 128:100–106, 2019.

Lyu, H. and Li, Y. Block majorization-minimization with diminishing radius for constrained nonconvex optimization. 08 2023.

Marcel, S. and Rodriguez, Y. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1485–1488, 2010.

Maron, H. and Lipman, Y. (probably) concave graph matching. *Advances in Neural Information Processing Systems*, 31, 2018.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Mémoli, F. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487, 2011.

Mémoli, F. and Needham, T. Gromov-monge quasi-metrics and distance distributions. *arXiv*, 2018, 2018.

Mémoli, F. and Needham, T. Comparison results for gromov-wasserstein and gromov-monge distances. *arXiv preprint arXiv:2212.14123*, 2022.

Nagano, Y., Yamaguchi, S., Fujita, Y., and Koyama, M. A wrapped normal distribution on hyperbolic space for gradient-based learning. In *International Conference on Machine Learning*, volume 97, pp. 4693–4702. PMLR, 09–15 Jun 2019.

Nene, S. A., Nayar, S. K., Murase, H., et al. Columbia object image library (coil-20). 1996.

Nickel, M. and Kiela, D. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, volume 80, pp. 3779–3788, 10–15 Jul 2018.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Peyré, G., Cuturi, M., and Solomon, J. Gromov-wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pp. 2664–2672. PMLR, 2016.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Redko, I., Vayer, T., Flamary, R., and Courty, N. Co-optimal transport. *Advances in Neural Information Processing Systems*, 33(17559-17570):2, 2020.

Rosenberg, A. and Hirschberg, J. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 410–420, 2007.

Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Roweis, S. T. and Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500): 2323–2326, 2000.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., and Lin, C.-T. A review of clustering techniques and developments. *Neurocomputing*, 267:664–681, 2017.

Scetbon, M., Peyré, G., and Cuturi, M. Linear-time gromov wasserstein distances using low rank couplings and costs. In *International Conference on Machine Learning*, pp. 19347–19365. PMLR, 2022.

Schaeffer, S. E. Graph clustering. *Computer science review*, 1(1):27–64, 2007.

Schölkopf, B., Smola, A., and Müller, K.-R. Kernel principal component analysis. In *International conference on artificial neural networks*, pp. 583–588. Springer, 1997.

Sinkhorn, R. and Knopp, P. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Solomon, J., Peyré, G., Kim, V. G., and Sra, S. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016.

Sturm, K.-T. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. *arXiv preprint arXiv:1208.0434*, 2012.

Tenenbaum, J. B., Silva, V. d., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Thual, A., TRAN, Q. H., Zemskova, T., Courty, N., Flamary, R., Dehaene, S., and Thirion, B. Aligning individual brains with fused unbalanced gromov wasserstein. *Advances in Neural Information Processing Systems*, 35: 21792–21804, 2022.

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

Van Assel, H., Espinasse, T., Chiquet, J., and Picard, F. A probabilistic graph coupling view of dimension reduction. *Advances in Neural Information Processing Systems*, 35: 10696–10708, 2022.

Van Assel, H., Vayer, T., Flamary, R., and Courty, N. SNEkhorn: Dimension reduction with symmetric entropic affinities. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Van Der Maaten, L., Postma, E., Van den Herik, J., et al. Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71), 2009.

Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. Optimal transport for structured data with application on graphs. *arXiv preprint arXiv:1805.09114*, 2018.

Ventre, E., Herbach, U., Espinasse, T., Benoit, G., and Gandrillon, O. One model fits all: combining inference and simulation of gene regulatory networks. *PLoS Computational Biology*, 19(3):e1010962, 2023.

Villani, C. et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Vincent-Cuaz, C. *Optimal transport for graph representation learning*. PhD thesis, Université Côte d'Azur, 2023.

Vincent-Cuaz, C., Vayer, T., Flamary, R., Corneli, M., and Courty, N. Online graph dictionary learning. In *International conference on machine learning*, pp. 10564–10574. PMLR, 2021.

Vincent-Cuaz, C., Flamary, R., Corneli, M., Vayer, T., and Courty, N. Semi-relaxed gromov-wasserstein divergence and applications on graphs. In *International Conference on Learning Representations*, 2022a.

Vincent-Cuaz, C., Flamary, R., Corneli, M., Vayer, T., and Courty, N. Template based graph neural network with optimal transport distances. *Advances in Neural Information Processing Systems*, 35:11800–11814, 2022b.

Vladymyrov, M. and Carreira-Perpinan, M. Entropic affinities: Properties and efficient numerical computation. In *International conference on machine learning*, pp. 477–485. PMLR, 2013.

Von Luxburg, U. A tutorial on spectral clustering. *Statistics and computing*, 17:395–416, 2007.

Wolf, F. A., Angerer, P., and Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

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, H., Luo, D., and Carin, L. Scalable gromov-wasserstein learning for graph partitioning and matching. *Advances in neural information processing systems*, 32, 2019.

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.

Zeng, Z., Zhu, R., Xia, Y., Zeng, H., and Tong, H. Generative graph dictionary learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40749–40769. PMLR, 23–29 Jul 2023.

A. Proof of results in Section 3

A.1. Proof of Lemma 3.1

We recall the result.

Lemma A.1. *Let* $\mathbf{C}_Z$ *be a permutation equivariant function and* L *any loss. The minimum eq. (DR) is equal to*

$$\min_{\mathbf{Z}\in\mathbb{R}^{N\times d}} \min_{\sigma\in S_N} \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{\sigma(i)\sigma(j)}) . \tag{2}$$

Also, any sol. $\mathbf{Z}$ *of eq. (DR) is such that* $(\mathbf{Z}, \mathrm{id})$ *is sol. of eq. (2) and conversely any* $(\mathbf{Z}, \sigma)$ *sol. of eq. (2) is such that* $\mathbf{Z}$ *is a sol of eq. (DR) up to* σ.

Proof. By suboptimality of $\sigma = \mathrm{id}$ we clearly have $\min_{\mathbf{Z}\in\mathbb{R}^{N\times d}} \min_{\sigma\in S_N} \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{\sigma(i)\sigma(j)}) \leq \min_{\mathbf{Z}\in\mathbb{R}^{N\times d}} \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{ij})$. For the other direction, take an optimal solution $(\mathbf{Z}, \sigma)$ of eq. (2). Using the permutation equivariance of $\mathbf{C}_Z$, $[\mathbf{C}_Z(\mathbf{Z})]_{\sigma(i)\sigma(j)} = [\mathbf{P}\mathbf{C}_Z(\mathbf{Z})\mathbf{P}^\top]_{ij} = [\mathbf{C}_Z(\mathbf{P}\mathbf{Z})]_{ij}$ for some permutation matrix $\mathbf{P}$. But $\mathbf{P}\mathbf{Z}$ is admissible for problem eq. (DR). Hence $\min_{\mathbf{Z}\in\mathbb{R}^{N\times d}} \min_{\sigma\in S_N} \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{\sigma(i)\sigma(j)}) \geq \min_{\mathbf{Z}\in\mathbb{R}^{N\times d}} \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{ij})$. $\square$

A.2. Proof of Theorem 3.2

In the following DS is the space of $N \times N$ doubly stochastic matrices. We begin by proving the first point of Theorem 3.2. We will rely on the simple, but useful, result below

Proposition A.2. *Let* $\Omega \subseteq \mathbb{R}$ *and* $\mathrm{Im}(\mathbf{C}_X) \subseteq \Omega^{N\times N}$. *Suppose that* $L(a, \cdot)$ *is convex for any* $a \in \Omega$ *and*

$$\min_{\mathbf{Z}\in\mathbb{R}^{N\times d}} \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{ij}) \leq \min_{\mathbf{Z}\in\mathbb{R}^{N\times d}, \mathbf{T}\in\mathrm{DS}} \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{T}\mathbf{C}_Z(\mathbf{Z})\mathbf{T}^\top]_{ij}) . \tag{4}$$

Then the minimum eq. (DR) is equal to $\min_{\mathbf{Z}} \mathrm{GW}_L(\mathbf{C}_X(\mathbf{X}), \mathbf{C}_Z(\mathbf{Z}), \frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N)$.

Proof. Consider any doubly stochastic matrix $\mathbf{T}$ and note that $[\mathbf{T}\mathbf{C}_Z(\mathbf{Z})\mathbf{T}^\top]_{ij} = \sum_{kl}[\mathbf{C}_Z(\mathbf{Z})]_{kl}T_{ik}T_{jl}$. Using the convexity of $L(a, \cdot)$ for any $a \in \Omega$ and Jensen's inequality we have

$$\sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{T}\mathbf{C}_Z(\mathbf{Z})\mathbf{T}^\top]_{ij}) = \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, \sum_{kl}[\mathbf{C}_Z(\mathbf{Z})]_{kl}T_{ik}T_{jl}) \leq \sum_{ijkl} L([\mathbf{C}_\mathbf{X}]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{kl})T_{ik}T_{jl} . \tag{5}$$

In particular

$$\min_{\mathbf{Z}} \min_{\mathbf{T}\in\mathrm{DS}} \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{T}\mathbf{C}_Z(\mathbf{Z})\mathbf{T}^\top]_{ij}) \leq \min_{\mathbf{Z}} \min_{\mathbf{T}\in\mathrm{DS}} \sum_{ijkl} L([\mathbf{C}_\mathbf{X}]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{kl})T_{ik}T_{jl} . \tag{6}$$

Hence, using eq. (4),

$$\min_{\mathbf{Z}} \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{ij}) \leq \min_{\mathbf{Z}} \min_{\mathbf{T}\in\mathrm{DS}} \sum_{ijkl} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{kl})T_{ik}T_{jl} . \tag{7}$$

But the converse inequality is also true by sub-optimality of $\mathbf{T} = \mathbf{I}_N$ for the problem $\min_{\mathbf{Z}} \min_{\mathbf{T}\in\mathrm{DS}} \sum_{ijkl} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{kl})T_{ik}T_{jl}$. Overall

$$\min_{\mathbf{Z}} \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{ij}) = \min_{\mathbf{Z}} \min_{\mathbf{T}\in\mathrm{DS}} \sum_{ijkl} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{kl})T_{ik}T_{jl} . \tag{8}$$

Now we conclude by using that the RHS of this equation is equivalent to the minimization in $\mathbf{Z}$ of $\mathrm{GW}_L(\mathbf{C}_X(\mathbf{X}), \mathbf{C}_Z(\mathbf{Z}), \frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N)$ (both problems only differ from a constant scaling factor N^2). $\square$

As a consequence we have the following result.

Proposition A.3. *Let* $\Omega \subseteq \mathbb{R}$ *and* $\mathrm{Im}(\mathbf{C}_X) \subseteq \Omega^{N \times N}$. *The minimum eq.* (DR) *is equal to* $\min_{\mathbf{Z}} \mathrm{GW}_L(\mathbf{C}_X(\mathbf{X}), \mathbf{C}_Z(\mathbf{Z}), \frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N)$ *when:*

(i) $L(a, \cdot)$ *is convex for any* $a \in \Omega$ *and the image of* $\mathbf{C}_Z$ *is stable by doubly-stochastic matrices, i.e. ,*

$$\forall \mathbf{Z} \in \mathbb{R}^{N \times d}, \forall \mathbf{T} \in \mathrm{DS}, \exists \mathbf{Y} \in \mathbb{R}^{N \times d}, \; \mathbf{C}_Z(\mathbf{Y}) = \mathbf{T}\mathbf{C}_Z(\mathbf{Z})\mathbf{T}^\top . \tag{9}$$

(ii) $L(a, \cdot)$ *is convex <u>and non-decreasing</u> for any* $a \in \Omega$ *and*

$$\forall \mathbf{Z} \in \mathbb{R}^{N \times d}, \forall \mathbf{T} \in \mathrm{DS}, \exists \mathbf{Y} \in \mathbb{R}^{N \times d}, \; \mathbf{C}_Z(\mathbf{Y}) \leq \mathbf{T}\mathbf{C}_Z(\mathbf{Z})\mathbf{T}^\top , \tag{10}$$

where $\leq$ *is understood element-wise, i.e. ,* $\mathbf{A} \leq \mathbf{B} \iff \forall(i,j), \; A_{ij} \leq B_{ij}$.

Proof. For the first point it suffices to see that the condition eq. (9) implies that $\{\mathbf{T}\mathbf{C}_Z(\mathbf{Z})\mathbf{T}^\top : \mathbf{Z} \in \mathbb{R}^{N \times d}, \mathbf{T} \in \mathrm{DS}\} \subseteq \{\mathbf{C}_Z(\mathbf{Z}) : \mathbf{Z} \in \mathbb{R}^{N \times d}\}$ (in fact we have equality by choosing $\mathbf{T} = \mathbf{I}_N$) and thus eq. (4) holds and we apply Proposition A.2.

For the second point we will also show that eq. (4) holds. Consider $\mathbf{Z}^\star, \mathbf{T}^\star$ minimizers of $\min_{\mathbf{Z}, \mathbf{T} \in \mathrm{DS}} \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{T}\mathbf{C}_Z(\mathbf{Z})\mathbf{T}^\top]_{ij})$. By hypothesis there exists $\mathbf{Y} \in \mathbb{R}^{N \times d}$ such that $\forall(i,j) \in [\![N]\!]^2, [\mathbf{T}^\star \mathbf{C}_Z(\mathbf{Z}^\star)\mathbf{T}^{\star\top}]_{ij} \geq [\mathbf{C}_Z(\mathbf{Y})]_{ij}$. Since $L([\mathbf{C}_X(\mathbf{X})]_{ij}, \cdot)$ is non-decreasing for any (i,j) then $\sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{T}^\star \mathbf{C}_Z(\mathbf{Z}^\star)\mathbf{T}^{\star\top}]_{ij}) \geq \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Y})]_{ij})$ and thus $\sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{T}^\star \mathbf{C}_Z(\mathbf{Z}^\star)\mathbf{T}^{\star\top}]_{ij}) \geq \min_{\mathbf{Z}} \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{ij})$ which gives the condition eq. (4) and we have the conclusion by Proposition A.2. $\qquad\square$

We recall that a function $R : \mathbb{R}^{N \times d} \to \mathbb{R}$ is called permutation invariant if $R(\mathbf{P}\mathbf{Z}) = R(\mathbf{Z})$ for any $\mathbf{Z} \in \mathbb{R}^{N \times d}$ and $N \times N$ permutation matrix $\mathbf{P}$. From the previous results we have the following corollary, which proves, in particular, the first point of Theorem 3.2.

Corollary A.4. *We have the following equivalences:*

(i) Consider the spectral methods which correspond to $\mathbf{C}_Z(\mathbf{Z}) = \mathbf{Z}\mathbf{Z}^\top$ *and* $L = L_2$. *Then for any* $\mathbf{C}_X$

$$\min_{\mathbf{Z} \in \mathbb{R}^{N \times d}} \sum_{ij} L_2([\mathbf{C}_\mathbf{X}]_{ij}, \langle \mathbf{z}_i, \mathbf{z}_j \rangle) \tag{11}$$

and

$$\min_{\mathbf{Z} \in \mathbb{R}^{N \times d}} \mathrm{GW}_{L_2}(\mathbf{C}_X(\mathbf{X}), \mathbf{Z}\mathbf{Z}^\top, \frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N) \tag{12}$$

are equal.

(ii) Consider the cross-entropy loss $L(x, y) = x \log(x/y)$ *and* $\mathbf{C}_X$ *such that* $\mathrm{Im}(\mathbf{C}_X) \subseteq \mathbb{R}_+^{N \times N}$. *Suppose that the similarity in the output space can be written as*

$$\forall(i,j) \in [\![N]\!]^2, [\mathbf{C}_Z(\mathbf{Z})]_{ij} = f(\mathbf{z}_i - \mathbf{z}_j)/R(\mathbf{Z}), \tag{13}$$

for some logarithmically concave function $f : \mathbb{R}^d \to \mathbb{R}_+$ *and normalizing factor* $R : \mathbb{R}^{N \times d} \to \mathbb{R}_+^*$ *which is both convex and permutation invariant. Then,*

$$\min_{\mathbf{Z} \in \mathbb{R}^{N \times d}} \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{ij}) \tag{14}$$

and

$$\min_{\mathbf{Z} \in \mathbb{R}^{N \times d}} \mathrm{GW}_L(\mathbf{C}_X(\mathbf{X}), \mathbf{C}_Z(\mathbf{Z}), \frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N) \tag{15}$$

are equal.

Proof. For the first point we show that the condition eq. (9) of Proposition A.3 is satisfied. Indeed take any $\mathbf{Z}, \mathbf{T}$ then $\mathbf{T}\mathbf{C}_Z(\mathbf{Z})\mathbf{T}^\top = \mathbf{T}\mathbf{Z}\mathbf{Z}^\top\mathbf{T}^\top = (\mathbf{T}\mathbf{Z})(\mathbf{T}\mathbf{Z})^\top = \mathbf{C}_Z(\mathbf{T}\mathbf{Z})$.

For the second point if we consider $\tilde{L}(a,b) = a \times b$ then we use that the neighbor embedding problem eq. (14) is equivalent to $\min_{\mathbf{Z}\in\mathbb{R}^{N\times d}} \sum_{ij} -[\mathbf{C}_X(\mathbf{X})]_{ij}\log([\mathbf{C}_Z(\mathbf{Z})]_{ij}) = \min_{\mathbf{Z}\in\mathbb{R}^{N\times d}} \sum_{ij} \tilde{L}([\mathbf{C}_X(\mathbf{X})]_{ij}, [\widetilde{\mathbf{C}_Z}(\mathbf{Z})]_{ij})$ where $[\widetilde{\mathbf{C}_Z}(\mathbf{Z})]_{ij} = g(\mathbf{z}_i - \mathbf{z}_j) + \log(R(\mathbf{Z}))$ with $g = -\log \circ f$. Since f is logarithmically concave g is convex. Moreover we have that $\tilde{L}(a, \cdot)$ is convex (it is linear) and non-decreasing since $a \in \mathbb{R}_+$ in this case ($\mathbf{C}_\mathbf{X}$ is non-negative). Also for any $\mathbf{Z} \in \mathbb{R}^{N\times d}$ and $\mathbf{T} \in \mathrm{DS}$ we have, using Jensen's inequality since $\mathbf{T}$ is doubly-stochastic,

$$
\begin{aligned}
[\mathbf{T}\widetilde{\mathbf{C}_Z}(\mathbf{Z})\mathbf{T}^\top]_{ij} &= \sum_{kl} g(\mathbf{z}_k - \mathbf{z}_l)T_{ik}T_{jl} + \log(R(\mathbf{Z})) \\
&\geq g(\sum_{kl}(\mathbf{z}_k - \mathbf{z}_l)T_{ik}T_{jl}) + \log(R(\mathbf{Z})) \\
&= g(\sum_k \mathbf{z}_k T_{ik} - \sum_l \mathbf{z}_l T_{jl}) + \log(R(\mathbf{Z})) \,.
\end{aligned}
\tag{16}
$$

Now we will prove that $\log(R(\mathbf{Z})) \geq \log(R(\mathbf{T}\mathbf{Z}))$. Using Birkhoff's theorem (Birkhoff, 1946) the matrix $\mathbf{T}$ can be decomposed as a convex combination of permutation matrices, *i.e.*, $\mathbf{T} = \sum_k \lambda_k \mathbf{P}_k$ where $(\mathbf{P}_k)_k$ are permutation matrices and $\lambda_k \in \mathbb{R}_+$ with $\sum_k \lambda_k = 1$. Hence by convexity and Jensen's inequality $R(\mathbf{T}\mathbf{Z}) = R(\sum_k \lambda_k \mathbf{P}_k \mathbf{Z}) \leq \sum_k \lambda_k R(\mathbf{P}_k \mathbf{Z})$. Now using that R is permutation invariant we get $R(\mathbf{P}_k \mathbf{Z}) = R(\mathbf{Z})$ and thus $R(\mathbf{T}\mathbf{Z}) \leq \sum_k \lambda_k R(\mathbf{Z}) = R(\mathbf{Z})$. Since the logarithm is non-decreasing we have $\log(R(\mathbf{Z})) \geq \log(R(\mathbf{T}\mathbf{Z}))$ and, overall,

$$
[\mathbf{T}\widetilde{\mathbf{C}_Z}(\mathbf{Z})\mathbf{T}^\top]_{ij} \geq g(\sum_k \mathbf{z}_k T_{ik} - \sum_l \mathbf{z}_l T_{jl}) + \log(R(\mathbf{T}\mathbf{Z})) = [\widetilde{\mathbf{C}_Z}(\mathbf{T}\mathbf{Z})]_{ij} \,.
\tag{17}
$$

Thus if we introduce $\mathbf{Y} = \mathbf{T}\mathbf{Z}$ we have $[\mathbf{T}\widetilde{\mathbf{C}_Z}(\mathbf{Z})\mathbf{T}^\top]_{ij} \geq [\widetilde{\mathbf{C}_Z}(\mathbf{Y})]_{ij}$ and eq. (10) is satisfied. Thus we can apply Proposition A.3 and state that $\min_{\mathbf{Z}\in\mathbb{R}^{N\times d}} \sum_{ij} \tilde{L}([\mathbf{C}_X(\mathbf{X})]_{ij}, [\widetilde{\mathbf{C}_Z}(\mathbf{Z})]_{ij})$ and $\min_{\mathbf{Z}\in\mathbb{R}^{N\times d}} \mathrm{GW}_{\tilde{L}}(\mathbf{C}_X(\mathbf{X}), \widetilde{\mathbf{C}_Z}(\mathbf{Z}), \frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N)$ are equivalent which concludes as $\min_{\mathbf{Z}\in\mathbb{R}^{N\times d}} \mathrm{GW}_{\tilde{L}}(\mathbf{C}_X(\mathbf{X}), \widetilde{\mathbf{C}_Z}(\mathbf{Z}), \frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N) = \min_{\mathbf{Z}\in\mathbb{R}^{N\times d}} \mathrm{GW}_L(\mathbf{C}_X(\mathbf{X}), \mathbf{C}_Z(\mathbf{Z}), \frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N)$. $\square$

It remains to prove the second point of Theorem 3.2 as stated below.

Proposition A.5. *Consider* $\mathrm{Im}(\mathbf{C}_X) \subseteq \mathbb{R}_+^{N\times N}$, $L = L_{\mathrm{KL}}$. *Suppose that for any* $\mathbf{X}$ *the matrix* $\mathbf{C}_X(\mathbf{X})$ *is CPD and for any* $\mathbf{Z}$

$$
\mathbf{C}_Z(\mathbf{Z}) = \mathrm{diag}(\boldsymbol{\alpha}_\mathbf{Z})\mathbf{K}_\mathbf{Z}\,\mathrm{diag}(\boldsymbol{\beta}_\mathbf{Z})\,,
\tag{18}
$$

where $\boldsymbol{\alpha}_\mathbf{Z}, \boldsymbol{\beta}_\mathbf{Z} \in \mathbb{R}_{>0}^N$ *and* $\mathbf{K}_\mathbf{Z} \in \mathbb{R}_{>0}^{N\times N}$ *is such that* $\log(\mathbf{K}_\mathbf{Z})$ *is CPD. Then the minimum eq.* (DR) *is equal to* $\min_\mathbf{Z} \mathrm{GW}_L(\mathbf{C}_X(\mathbf{X}), \mathbf{C}_Z(\mathbf{Z}), \frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N)$.

Proof. To prove this result we will show that, for any $\mathbf{Z}$, the function

$$
\mathbf{T} \in \mathcal{U}(\frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N) \to E_L(\mathbf{C}_X(\mathbf{X}), \mathbf{C}_Z(\mathbf{Z}), \mathbf{T})\,,
\tag{19}
$$

is actually concave. Indeed, in this case there exists a minimizer which is an extremal point of $\mathcal{U}(\frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N)$. By Birkhoff's theorem (Birkhoff, 1946) these extreme points are the matrices $\frac{1}{N}\mathbf{P}$ where $\mathbf{P}$ is a $N \times N$ permutation matrix. Consequently, when the function eq. (19) is concave minimizing $\mathrm{GW}_L(\mathbf{C}_X(\mathbf{X}), \mathbf{C}_Z(\mathbf{Z}), \frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N)$ in $\mathbf{Z}$ is equivalent to minimizing in $\mathbf{Z}$

$$
\min_{\mathbf{P}\in\mathbb{R}^{N\times N}\text{ permutation}} \sum_{ijkl} L([\mathbf{C}_X(\mathbf{X})]_{ik}, [\mathbf{C}_Z(\mathbf{Z})]_{jl})P_{ij}P_{kl} = \min_{\sigma\in S_N} \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{\sigma(i)\sigma(j)})\,,
\tag{20}
$$

which is exactly the Gromov-Monge problem described in Lemma 3.1 and thus the problem is equivalent to eq. (DR) by Lemma 3.1.

15

First note that $L(x,y) = x\log(x) - x - x\log(y) + y$ so for any $\mathbf{T} \in \mathcal{U}(\frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N)$ the loss $E_L(\mathbf{C}_X(\mathbf{X}), \mathbf{C}_Z(\mathbf{Z}), \mathbf{T})$ is equal to

$$
\begin{aligned}
\sum_{ijkl} L([\mathbf{C}_X(\mathbf{X})]_{ik}, [\mathbf{C}_Z(\mathbf{Z})]_{jl}) T_{ij} T_{kl} &= a_{\mathbf{X}} + b_{\mathbf{Z}} - \sum_{ijkl} [\mathbf{C}_X(\mathbf{X})]_{ik} \log([\mathbf{C}_Z(\mathbf{Z})]_{jl}) T_{ij} T_{kl} \\
&= a_{\mathbf{X}} + b_{\mathbf{Z}} - \sum_{ijkl} [\mathbf{C}_X(\mathbf{X})]_{ik} \log([\boldsymbol{\alpha}_{\mathbf{Z}}]_j [\boldsymbol{\beta}_{\mathbf{Z}}]_l [\mathbf{W}]_{jl}) T_{ij} T_{kl} \\
&= a_{\mathbf{X}} + b_{\mathbf{Z}} - \frac{1}{N} \sum_{ijk} [\mathbf{C}_X(\mathbf{X})]_{ik} \log([\boldsymbol{\alpha}_{\mathbf{Z}}]_j) T_{ij} \\
&\quad - \frac{1}{N} \sum_{ikl} [\mathbf{C}_X(\mathbf{X})]_{ik} \log([\boldsymbol{\beta}_{\mathbf{Z}}]_l) T_{kl} - \sum_{ijkl} [\mathbf{C}_X(\mathbf{X})]_{ik} \log([\mathbf{K}_{\mathbf{Z}}]_{jl}) T_{ij} T_{kl} \,,
\end{aligned}
\tag{21}
$$

where $a_{\mathbf{X}}, b_{\mathbf{Z}}$ are terms that do not depend on $\mathbf{T}$. Since the problem is quadratic the concavity only depends on the term $-\sum_{ijkl}[\mathbf{C}_X(\mathbf{X})]_{ik}\log([\mathbf{K}_{\mathbf{Z}}]_{jl})T_{ij}T_{kl} = -\operatorname{Tr}(\mathbf{C}_X(\mathbf{X})\mathbf{T}^\top \log(\mathbf{K}_{\mathbf{Z}})\mathbf{T})$. From Maron & Lipman (2018) we know that the function $\mathbf{T} \to -\operatorname{Tr}(\mathbf{C}_X(\mathbf{X})\mathbf{T}^\top \log(\mathbf{K}_{\mathbf{Z}})\mathbf{T})$ is concave on $\mathcal{U}(\frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N)$ when $\mathbf{C}_X(\mathbf{X})$ is CPD and $\log(\mathbf{W}_{\mathbf{Z}})$ is CPD. This concludes the proof.

$\square$

A.3. Necessary and sufficient condition

We give a necessary and sufficient condition under which the DR problem is equivalent to a GW problem

Proposition A.6. *Let $\mathbf{C}_1 \in \mathbb{R}^{N \times N}, L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and $\mathcal{C} \subseteq \mathbb{R}^{N \times N}$ a subspace of $N \times N$ matrices. We suppose that $\mathcal{C}$ is stable by permutations i.e. , $\mathbf{C} \in \mathcal{C}$ implies that $\mathbf{P}\mathbf{C}\mathbf{P}^\top \in \mathcal{C}$ for any $N \times N$ permutation matrix $\mathbf{P}$. Then*

$$
\min_{\mathbf{C}_2 \in \mathcal{C}} \sum_{(i,j) \in [\![N]\!]^2} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{ij}) = \min_{\mathbf{C}_2 \in \mathcal{C}} \min_{\substack{\mathbf{T} \in \mathbb{R}_+^{N \times N} \\ \mathbf{T}\mathbf{1}_N = \mathbf{1}_N \\ \mathbf{T}^\top \mathbf{1}_N = \mathbf{1}_N}} \sum_{ijkl} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{kl}) T_{ik} T_{jl}
\tag{22}
$$

if and only if the optimal assignment problem

$$
\min_{\sigma_1, \sigma_2 \in S_N} f(\sigma_1, \sigma_2) := \min_{\mathbf{C}_2 \in \mathcal{C}} \sum_{ij} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{\sigma_1(i)\sigma_2(j)})
\tag{23}
$$

admits an optimal solution $(\sigma_1^\star, \sigma_2^\star)$ with $\sigma_1^\star = \sigma_2^\star$.

Proof. We note that the LHS of eq. (22) is always smaller than the RHS since $\mathbf{T} = \mathbf{I}_N$ is admissible for the RHS problem. So both problems are equal if and only if

$$
\min_{\mathbf{C}_2 \in \mathcal{C}} \sum_{(i,j) \in [\![N]\!]^2} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{ij}) \leq \min_{\mathbf{C}_2 \in \mathcal{C}} \min_{\substack{\mathbf{T} \in \mathbb{R}_+^{N \times N} \\ \mathbf{T}\mathbf{1}_N = \mathbf{1}_N \\ \mathbf{T}^\top \mathbf{1}_N = \mathbf{1}_N}} \sum_{ijkl} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{kl}) T_{ik} T_{jl} \,.
\tag{24}
$$

Now consider any $\mathbf{C}_2$ fixed and observe that

$$
\min_{\mathbf{T} \in \mathrm{DS}} \sum_{ijkl} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{kl}) T_{ik} T_{jl} \geq \min_{\mathbf{T}^{(1)}, \mathbf{T}^{(2)} \in \mathrm{DS}} \sum_{ijkl} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{kl}) T_{ik}^{(1)} T_{jl}^{(2)} \,.
\tag{25}
$$

The latter problem is a co-optimal transport problem (Redko et al., 2020), and, since it is a bilinear problem, there is an optimal solution $(\mathbf{T}^{(1)}, \mathbf{T}^{(2)})$ such that both $\mathbf{T}^{(1)}$ and $\mathbf{T}^{(2)}$ are in an extremal point of DS which is the space of $N \times N$ permutation matrices by Birkhoff's theorem (Birkhoff, 1946). This point was already noted by (Konno, 1976) but we recall the proof for completeness. We note $L_{ijkl} = L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{kl})$ and consider an optimal solution $(\mathbf{T}_\star^{(1)}, \mathbf{T}_\star^{(2)})$ of $\min_{\mathbf{T}^{(1)}, \mathbf{T}^{(2)} \in \mathrm{DS}} \phi(\mathbf{T}^{(1)}, \mathbf{T}^{(2)}) := \sum_{ijkl} L_{ijkl} T_{ik}^{(1)} T_{jl}^{(2)}$. Consider the linear problem $\min_{\mathbf{T} \in \mathrm{DS}} \phi(\mathbf{T}, \mathbf{T}_\star^{(2)})$. Since it is a linear over the space of doubly stochastic matrices it admits a permutation matrix $\mathbf{P}^{(1)}$ as optimal solution. Also

$\phi(\mathbf{P}^{(1)}, \mathbf{T}_\star^{(2)}) \leq \phi(\mathbf{T}_\star^{(1)}, \mathbf{T}_\star^{(2)})$ by optimality. Now consider the linear problem $\min_{\mathbf{T} \in \mathrm{DS}} \phi(\mathbf{P}^{(1)}, \mathbf{T})$, in the same it admits a permutation matrix $\mathbf{P}^{(2)}$ as optimal solution and by optimality $\phi(\mathbf{P}^{(1)}, \mathbf{P}^{(2)}) \leq \phi(\mathbf{P}^{(1)}, \mathbf{T}_\star^{(2)})$ thus $\phi(\mathbf{P}^{(1)}, \mathbf{P}^{(2)}) \leq \phi(\mathbf{T}_\star^{(1)}, \mathbf{T}_\star^{(2)})$ which implies that $(\mathbf{P}^{(1)}, \mathbf{P}^{(2)})$ is an optimal solution. Combining with eq. (25) we get

$$
\min_{\mathbf{C}_2 \in \mathcal{C}} \min_{\mathbf{T} \in \mathrm{DS}} \sum_{ijkl} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{kl}) \, T_{ik} T_{jl} \geq \min_{\mathbf{C}_2 \in \mathcal{C}} \min_{\sigma_1, \sigma_2 \in S_N} \sum_{ij} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{\sigma_1(i)\sigma_2(j)})
$$
$$
= \min_{\sigma_1, \sigma_2 \in S_N} f(\sigma_1, \sigma_2). \tag{26}
$$

Now suppose that the optimal assignment problem $\min_{\sigma_1, \sigma_2 \in S_N} f(\sigma_1, \sigma_2)$ admits an optimal solution $(\sigma_1^\star, \sigma_2^\star)$ with $\sigma_1^\star = \sigma_2^\star$. Then with eq. (29)

$$
\min_{\mathbf{C}_2 \in \mathcal{C}} \min_{\mathbf{T} \in \mathrm{DS}} \sum_{ijkl} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{kl}) \, T_{ik} T_{jl} \geq \min_{\mathbf{C}_2 \in \mathcal{C}} \sum_{ij} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{\sigma_1^\star(i)\sigma_1^\star(j)})
$$
$$
\geq \min_{\sigma \in S_N} \min_{\mathbf{C}_2 \in \mathcal{C}} \sum_{ij} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{\sigma(i)\sigma(j)}). \tag{27}
$$

Now since $\mathcal{C}$ is stable by permutation then $\{([\mathbf{C}]_{\sigma(i)\sigma(j)})_{(i,j) \in [\![N]\!]^2} : \mathbf{C} \in \mathcal{C}, \sigma \in S_N\} = \mathcal{C}$ and consequently $\min_{\sigma \in S_N} \min_{\mathbf{C}_2 \in \mathcal{C}} \sum_{ij} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{\sigma(i)\sigma(j)}) = \min_{\mathbf{C}_2 \in \mathcal{C}} \sum_{ij} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{ij})$. Consequently, using eq. (27),

$$
\min_{\mathbf{C}_2 \in \mathcal{C}} \min_{\mathbf{T} \in \mathrm{DS}} \sum_{ijkl} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{kl}) \, T_{ik} T_{jl} \geq \min_{\mathbf{C}_2 \in \mathcal{C}} \sum_{ij} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{ij}), \tag{28}
$$

and thus both are equal.

Conversely suppose that eq. (22) holds. Then, from eq. (29) we have

$$
\min_{\sigma_1, \sigma_2 \in S_N} f(\sigma_1, \sigma_2) = \min_{\sigma_1, \sigma_2 \in S_N} \min_{\mathbf{C}_2 \in \mathcal{C}} \sum_{ij} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{\sigma_1(i)\sigma_2(j)}) \leq \min_{\mathbf{C}_2 \in \mathcal{C}} \min_{\mathbf{T} \in \mathrm{DS}} \sum_{ijkl} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{kl}) \, T_{ik} T_{jl}
$$
$$
= \min_{\mathbf{C}_2 \in \mathcal{C}} \sum_{(i,j) \in [\![N]\!]^2} L([\mathbf{C}_1]_{ij}, [\mathbf{C}_2]_{ij}) = f(\mathrm{id}, \mathrm{id}), \tag{29}
$$

which concludes the proof. □

Remark A.7. The condition on the set of similarity matrices $\mathcal{C}$ is quite reasonable: it indicates that if $\mathbf{C}$ is an admissible similarity matrix, then permuting the rows and columns of $\mathbf{C}$ results in another admissible similarity matrix. For DR, the corresponding $\mathcal{C}$ is $\mathcal{C} = \{\mathbf{C}_Z(\mathbf{Z}) : \mathbf{Z} \in \mathbb{R}^{N \times d}\}$. In this case, if $\mathbf{C}_Z(\mathbf{Z})$ is of the form

$$
[\mathbf{C}_Z(\mathbf{Z})]_{ij} = h(f(\boldsymbol{z}_i, \boldsymbol{z}_j), g(\mathbf{Z})), \tag{30}
$$

where $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $h : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R}^{N \times d} \to \mathbb{R}$ which is permutation invariant (Bronstein et al., 2021), then $\mathcal{C}$ is stable under permutation. Indeed, permuting the rows and columns of $\mathbf{C}_Z(\mathbf{Z})$ by σ is equivalent to considering the similarity $\mathbf{C}_Z(\mathbf{Y})$, where $\mathbf{Y} = (\boldsymbol{z}_{\sigma(1)}, \cdots, \boldsymbol{z}_{\sigma(n)})^\top$. Moreover, most similarities in the target space considered in DR take the form eq. (30): $\langle \Phi(\boldsymbol{z}_i), \Phi(\boldsymbol{z}_j) \rangle_{\mathcal{H}}$ (kernels such as in spectral methods with $\Phi = \mathrm{id}$), $f(\boldsymbol{z}_i, \boldsymbol{z}_j)/\sum_{nm} f(\boldsymbol{z}_n, \boldsymbol{z}_m)$ (normalized similarities such as in SNE and t-SNE). Also note that the condition on $\mathcal{C} = \{\mathbf{C}_Z(\mathbf{Z}) : \mathbf{Z} \in \mathbb{R}^{N \times d}\}$ of Proposition A.6 is met as soon as $\mathbf{C}_Z : \mathbb{R}^{N \times d} \to \mathbb{R}^{N \times N}$ is permutation equivariant.

B. Generalized Semi-relaxed Gromov-Wasserstein is a divergence

Remark B.1 (Weak isomorphism). According to the notion of weak isomorphism in Chowdhury & Mémoli (2019), for a graph $(\mathbf{C}, \mathbf{h})$ with corresponding discrete measure $\mu = \sum_i h_i \delta_{x_i}$, two nodes x_i and x_j are the "same" if they have the same internal perception *i.e* $C_{ii} = C_{jj} = C_{ij} = C_{ji}$ and external perception $\forall k \neq (i,j), C_{ik} = C_{jk}, C_{ki} = C_{kj}$. So two graphs $(\mathbf{C}_1, \mathbf{h}_1)$ and $(\mathbf{C}_2, \mathbf{h}_2)$ are said to be weakly isomorphic, if there exist a canonical representation $(\mathbf{C}_c, \mathbf{h}_c)$ such that $\mathrm{card}(\mathrm{supp}(\mathbf{h}_c)) = p \leq n, m$ and $M_1 \in \{0,1\}^{n \times p}$ (resp. $M_2 \in \{0,1\}^{m \times p}$) such that for $k \in \{1, 2\}$

$$
\mathbf{C}_c = \mathbf{M}_k^\top \mathbf{C}_c \mathbf{M}_k \quad \text{and} \quad \mathbf{h}_c = \mathbf{M}_k^\top \mathbf{h}_k \tag{31}
$$

We first emphasize a simple result.

Proposition B.2. *Let any divergence* $L : \Omega \times \Omega \to \mathbb{R}_+$ *for* $\Omega \subseteq \mathbb{R}$, *then for any* $(\mathbf{C}, \mathbf{h})$ *and* $(\overline{\mathbf{C}}, \overline{\mathbf{h}})$, *we have* $\mathrm{GW}_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{h}, \overline{\mathbf{h}}) = 0$ *if and only if* $\mathrm{GW}_{L_2}(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{h}, \overline{\mathbf{h}}) = 0$.

Proof. If $\mathrm{GW}_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{h}, \overline{\mathbf{h}}) = 0$, then there exists $\mathbf{T} \in \mathcal{U}(\mathbf{h}, \overline{\mathbf{h}})$ such that

$$E_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{T}) = \sum_{ijkl} L(C_{ij}, \overline{C}_{kl}) T_{ik} T_{jl} = 0 \tag{32}$$

so whenever $T_{ik} T_{jl} \neq 0$, we must have $L(C_{ij}, \overline{C}_{kl}) = 0$ *i.e* $C_{ij} = \overline{C}_{kl}$ as L is a divergence. Which implies that $E_{L'}(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{T}) = 0$ for any other divergence L' well defined on any domain $\Omega \times \Omega$, necessarily including L_2. $\qquad\square$

Lemma B.3. *Let any divergence* $L : \Omega \times \Omega \to \mathbb{R}_+$ *for* $\Omega \subseteq \mathbb{R}$. *Let* $(\mathbf{C}, \mathbf{h}) \in \Omega^{n \times n} \times \Sigma_n$ *and* $(\overline{\mathbf{C}}, \overline{\mathbf{h}}) \in \mathbb{R}^{m \times m} \times \Sigma_m$. *Then* $\mathrm{srGW}_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{h}, \overline{\mathbf{h}}) = 0$ *if and only if there exists* $\overline{\mathbf{h}} \in \Sigma_m$ *such that* $(\mathbf{C}, \mathbf{h})$ *and* $(\overline{\mathbf{C}}, \overline{\mathbf{h}})$ *are weakly isomorphic.*

Proof. ($\Rightarrow$) As $\mathrm{srGW}_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{h}, \overline{\mathbf{h}}) = 0$ there exists $\mathbf{T} \in \mathcal{U}(\mathbf{h}, \overline{\mathbf{h}})$ such that $E_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{T}) = 0$ hence $\mathrm{GW}_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{h}, \overline{\mathbf{h}}) = 0$. Using proposition B.2, $\mathrm{GW}_{L_2} = 0$ hence using Theorem 18 in Chowdhury & Mémoli (2019), it implies that $(\mathbf{C}, \mathbf{h})$ and $(\overline{\mathbf{C}}, \overline{\mathbf{h}})$ are weakly isomorphic.

($\Leftarrow$) As mentioned, $(\mathbf{C}, \mathbf{h})$ and $(\overline{\mathbf{C}}, \overline{\mathbf{h}})$ being weakly isomorphic implies that $\mathrm{GW}_{L_2} = 0$. So there exists $\mathbf{T} \in \mathcal{U}(\mathbf{h}, \overline{\mathbf{h}})$, such that $E_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{T}) = 0$. Moreover $\mathbf{T}$ is admissible for the srGW problem as $\mathbf{T} \in \mathcal{U}(\mathbf{h}, \overline{\mathbf{h}}) \subset \mathcal{U}_n(\mathbf{h})$, thus $\mathrm{srGW}_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{h}, \overline{\mathbf{h}}) = 0$. $\qquad\square$

B.1. About trivial solutions of semi-relaxed GW when L is not a proper divergence

We briefly describe here some trivial solutions of srGW_L when L is not a proper divergence. We recall that

$$\mathrm{srGW}_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{h}) = \min_{\mathbf{T} \in \mathbb{R}_+^{N \times n} : \mathbf{T}\mathbf{1}_n = \mathbf{h}} E_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{T}) = \sum_{ijkl} L(C_{ij}, \overline{C}_{kl}) T_{ik} T_{jl}. \tag{33}$$

Suppose that $\mathbf{h} \in \Sigma_N^*$ and that $\overline{\mathbf{C}}$ has a minimum value on its diagonal *i.e.* $\min_{(i,j)} \overline{C}_{ij} = \min_{ii} \overline{C}_{ii}$. Suppose also that $\forall a, L(a, \cdot)$ is both convex *and* non-decreasing. First we have $\sum_j \frac{T_{ij}}{h_i} = 1$ for any $i \in [\![N]\!]$. Hence using the convexity of L, Jensen inequality and the fact that $L(a, \cdot)$ is non-decreasing for any a

$$\begin{aligned}
\sum_{ijkl} L(C_{ij}, \overline{C}_{kl}) T_{ik} T_{jl} &= \sum_{ijkl} L(C_{ij}, \overline{C}_{kl}) h_i h_j \frac{T_{ik}}{h_i} \frac{T_{jl}}{h_j} \\
&\geq \sum_{ij} L(C_{ij}, \sum_{kl} \overline{C}_{kl} \frac{T_{ik}}{h_i} \frac{T_{jl}}{h_j}) h_i h_j \\
&\geq \sum_{ij} L(C_{ij}, \sum_{kl} (\min_{nm} \overline{C}_{nm}) \frac{T_{ik}}{h_i} \frac{T_{jl}}{h_j}) h_i h_j \\
&= \sum_{ij} L(C_{ij}, (\min_{nm} \overline{C}_{nm})) h_i h_j \\
&= \sum_{ij} L(C_{ij}, (\min_{nn} \overline{C}_{nn})) h_i h_j.
\end{aligned} \tag{34}$$

Now suppose without loss of generality that $\min_{ii} \overline{C}_{ii} = \overline{C}_{11}$ then this gives

$$\min_{\mathbf{T} \in \mathbb{R}_+^{N \times n} : \mathbf{T}\mathbf{1}_n = \mathbf{h}} \sum_{ijkl} L(C_{ij}, \overline{C}_{kl}) T_{ik} T_{jl} \geq \sum_{ij} L(C_{ij}, \overline{C}_{11}) h_i h_j. \tag{35}$$

Now consider the coupling $\mathbf{T}^\star = \begin{pmatrix} h_1 & 0 & 0 & 0 \\ h_2 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ h_N & 0 & 0 & 0 \end{pmatrix}$. It is admissible and satisfies

$$E_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{T}^\star) = \sum_{ij} L(C_{ij}, \overline{C}_{11}) h_i h_j \leq \min_{\mathbf{T} \in \mathbb{R}_+^{N \times n} : \mathbf{T1}_n = \mathbf{h}} E_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{T}) \,. \tag{36}$$

Consequently the coupling $\mathbf{T}^\star$ is optimal. However the solution given by this coupling is trivial: it consists in sending all the mass to one unique point. In another words, all the nodes in the input graph are sent to a unique node in the target graph. Note that this phenomena is impossible for standard GW because of the coupling constraints.

We emphasize that this hypothesis on L *cannot be satisfied* as soon as L is a proper divergence. Indeed when L is a divergence the constraint "$L(a, \cdot)$ is non-decreasing for any a" is not possible as it would break the divergence constraints $\forall a, b\ L(a, b) \geq 0$ and $L(a, b) = 0 \iff a = b$ (at some point L must be decreasing).

C. Proof of Theorem 4.1

We recall that a matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$ is conditionally positive definite (CPD), *resp.* negative definite (CND), if $\forall \mathbf{x} \in \mathbb{R}^N, \mathbf{x}^\top \mathbf{1}_N = 0$ s.t. $\mathbf{x}^\top \mathbf{C} \mathbf{x} \geq 0$, *resp.* ≤ 0. We also consider the Hadamard product of matrices as $\mathbf{A} \odot \mathbf{B} = (A_{ij} \times B_{ij})_{ij}$. The i-th column of a matrix $\mathbf{T}$ is the vector denoted by $\mathbf{T}_{:,i}$. For a vector $\mathbf{x} \in \mathbb{R}^n$ we denote by $\mathrm{diag}(\mathbf{x})$ the diagonal $n \times n$ matrix whose elements are the x_i.

We state below the theorem that we prove in this section.

Theorem 4.1. *Let $\mathbf{h}_X \in \Sigma_N$ and $L = L_2$. Suppose that for any $\mathbf{X} \in \mathbb{R}^{N \times p}$ the matrix $\mathbf{C}_X(\mathbf{X})$ is CPD or CND. Then the problem eq. (srGWB) admits a membership matrix as optimal coupling, i.e. , there is a minimizer of $\mathbf{T} \in \mathcal{U}_n(\mathbf{h}_X) \to \min_{\overline{\mathbf{C}} \in \mathbb{R}^{n \times n}} E_L(\mathbf{C}_X(\mathbf{X}), \overline{\mathbf{C}}, \mathbf{T})$ with only one non-zero value per row.*

In order to prove this result we introduce the space of semi-relaxed couplings whose columns are not zero

$$\mathcal{U}_n^+(\mathbf{h}_X) = \{\mathbf{T} \in \mathcal{U}_n(\mathbf{h}_X) : \forall i \in [\![n]\!], \mathbf{T}_{:,i} \neq 0\} \,, \tag{37}$$

and we will use the following lemma.

Lemma C.1. *Let $\mathbf{h}_X \in \Sigma_N$, $L = L_2$ and $\mathbf{C}_X(\mathbf{X})$ symmetric. For any $\mathbf{T} \in \mathcal{U}_n(\mathbf{h}_X)$, the matrix $\overline{\mathbf{C}}(\mathbf{T}) \in \mathbb{R}^{n \times n}$ defined by*

$$\overline{\mathbf{C}}(\mathbf{T}) = \begin{cases} \mathbf{T}_{:,i}^\top \mathbf{C}_X(\mathbf{X}) \mathbf{T}_{:,j} / (\mathbf{T}_{:,i}^\top \mathbf{1}_N)(\mathbf{T}_{:,j}^\top \mathbf{1}_N) & \text{for } (i,j) \text{ such that } \mathbf{T}_{:,i} \text{ and } \mathbf{T}_{:,j} \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{38}$$

is a minimizer of $G : \overline{\mathbf{C}} \in \mathbb{R}^{n \times n} \to E_L(\mathbf{C}_X(\mathbf{X}), \overline{\mathbf{C}}, \mathbf{T})$. For $\mathbf{T} \in \mathcal{U}_n(\mathbf{h}_X)$ the expression of the minimum is

$$G(\mathbf{T}) = \mathrm{cte} - \mathrm{Tr}\left((\mathbf{T}^\top \mathbf{1}_N)(\mathbf{T}^\top \mathbf{1}_N)^\top (\overline{\mathbf{C}}(\mathbf{T}) \odot \overline{\mathbf{C}}(\mathbf{T}))\right) \,, \tag{39}$$

which defines a continuous function on $\mathcal{U}_n(\mathbf{h}_X)$. If $\mathbf{T} \in \mathcal{U}_n^+(\mathbf{h}_X)$ it becomes

$$G(\mathbf{T}) = \mathrm{cte} - \sum_{ij} \frac{(\mathbf{T}_{:,i}^\top \mathbf{C}_X(\mathbf{X}) \mathbf{T}_{:,j})^2}{(\mathbf{T}_{:,i}^\top \mathbf{1}_N)(\mathbf{T}_{:,j}^\top \mathbf{1}_N)} = \mathrm{cte} - \| \mathrm{diag}(\mathbf{T}^\top \mathbf{1}_N)^{-\frac{1}{2}} \mathbf{T}^\top \mathbf{C}_X(\mathbf{X}) \mathbf{T} \, \mathrm{diag}(\mathbf{T}^\top \mathbf{1}_N)^{-\frac{1}{2}} \|_F^2 \,. \tag{40}$$

Proof. First see that $\overline{\mathbf{C}}(\mathbf{T})$ is well defined since $\mathbf{T}_{:,i} \neq 0 \iff \mathbf{T}_{:,i}^\top \mathbf{1}_N \neq 0$ because $\mathbf{T}$ is non-negative. Consider, for $\mathbf{T} \in \mathcal{U}_n(\mathbf{h}_X)$, the function

$$F(\mathbf{T}, \overline{\mathbf{C}}) := E_L(\mathbf{C}_X(\mathbf{X}), \overline{\mathbf{C}}, \mathbf{T}) = \sum_{ijkl} ([\mathbf{C}_X(\mathbf{X})]_{ik} - \overline{C}_{jl})^2 T_{ij} T_{kl} \,, \tag{41}$$

A development yields (using $\mathbf{T}^\top \mathbf{1}_N = \mathbf{h}_X$)

$$F(\mathbf{T}, \overline{\mathbf{C}}) = \sum_{ik} [\mathbf{C}_X(\mathbf{X})]_{ik}^2 [\mathbf{h}_X]_i [\mathbf{h}_X]_k + \sum_{jl} \overline{C}_{jl}^2 (\sum_i T_{ij})(\sum_k T_{kl}) - 2 \sum_{ijkl} \overline{C}_{jl} [\mathbf{C}_X(\mathbf{X})]_{ik} T_{ij} T_{kl} \,. \tag{42}$$

We can rewrite $\sum_{ijkl} \overline{C}_{jl}[\mathbf{C}_X(\mathbf{X})]_{ik}T_{ij}T_{kl} = \mathrm{Tr}(\mathbf{T}^\top \mathbf{C}_X(\mathbf{X})\mathbf{T}\overline{\mathbf{C}})$. Also we have $(\sum_i T_{ij})(\sum_k T_{kl}) = [(\mathbf{T}^\top \mathbf{1}_N)(\mathbf{T}^\top \mathbf{1}_N)^\top]_{jl}$ Thus

$$\sum_{jl} \overline{C}_{jl}^2 (\sum_i T_{ij})(\sum_k T_{kl}) = \mathrm{Tr}((\mathbf{T}^\top \mathbf{1}_N)(\mathbf{T}^\top \mathbf{1}_N)^\top (\overline{\mathbf{C}} \odot \overline{\mathbf{C}})). \tag{43}$$

Overall

$$F(\mathbf{T}, \overline{\mathbf{C}}) = \mathrm{cte} + \mathrm{Tr}((\mathbf{T}^\top \mathbf{1}_N)(\mathbf{T}^\top \mathbf{1}_N)^\top (\overline{\mathbf{C}} \odot \overline{\mathbf{C}})) - 2\,\mathrm{Tr}(\mathbf{T}^\top \mathbf{C}_X(\mathbf{X})\mathbf{T}\overline{\mathbf{C}}). \tag{44}$$

Now taking the derivative with respect to $\overline{\mathbf{C}}$, the first order conditions are

$$\partial_2 F(\mathbf{T}, \overline{\mathbf{C}}) = 2(\overline{\mathbf{C}} \odot (\mathbf{T}^\top \mathbf{1}_N)(\mathbf{T}^\top \mathbf{1}_N)^\top - \mathbf{T}^\top \mathbf{C}_X(\mathbf{X})\mathbf{T}) = 0. \tag{45}$$

For (i,j) such that $\mathbf{T}_{:,i}$ and $\mathbf{T}_{:,j} \neq 0$ we have $[\partial_2 F(\mathbf{T}, \overline{\mathbf{C}}(\mathbf{T}))]_{ij} = 0$. For (i,j) such that $\mathbf{T}_{:,i}$ or $\mathbf{T}_{:,j} = 0$ we have $[\overline{\mathbf{C}} \odot (\mathbf{T}^\top \mathbf{1}_N)(\mathbf{T}^\top \mathbf{1}_N)^\top]_{ij} = 0$ and also $[\mathbf{T}^\top \mathbf{C}_X(\mathbf{X})\mathbf{T}]_{ij} = \mathbf{T}_{:,i}^\top \mathbf{C}_X(\mathbf{X})\mathbf{T}_{:,j} = 0$. In particular the matrix $\overline{\mathbf{C}}(\mathbf{T})$ satisfies the first order conditions. When $L = L_2$ the problem $\min_{\overline{\mathbf{C}} \in \mathbb{R}^{n \times n}} E_L(\mathbf{C}_X(\mathbf{X}), \overline{\mathbf{C}}, \mathbf{T}) = \frac{1}{2}\sum_{ijkl}([\mathbf{C}_X(\mathbf{X})]_{ik} - \overline{C}_{jl})^2 T_{ij}T_{kl}$ is convex in $\overline{\mathbf{C}}$. The first order conditions are sufficient hence $\overline{\mathbf{C}}(\mathbf{T})$ is a minimizer.

Also $\mathbf{T}^\top \mathbf{C}_X(\mathbf{X})\mathbf{T} = \overline{\mathbf{C}}(\mathbf{T}) \odot (\mathbf{T}^\top \mathbf{1}_N)(\mathbf{T}^\top \mathbf{1}_N)^\top$ by definition of $\overline{\mathbf{C}}(\mathbf{T})$ thus

$$\begin{aligned}
\mathrm{Tr}(\mathbf{T}^\top \mathbf{C}_X(\mathbf{X})\mathbf{T}\overline{\mathbf{C}}(\mathbf{T})) &= \mathrm{Tr}([\overline{\mathbf{C}}(\mathbf{T}) \odot (\mathbf{T}^\top \mathbf{1}_N)(\mathbf{T}^\top \mathbf{1}_N)^\top]\overline{\mathbf{C}}(\mathbf{T})) \\
&= \mathrm{Tr}((\mathbf{T}^\top \mathbf{1}_N)(\mathbf{T}^\top \mathbf{1}_N)^\top [\overline{\mathbf{C}}(\mathbf{T}) \odot \overline{\mathbf{C}}(\mathbf{T})]).
\end{aligned} \tag{46}$$

Hence

$$F(\mathbf{T}, \overline{\mathbf{C}}(\mathbf{T})) = \mathrm{cte} - \mathrm{Tr}\left((\mathbf{T}^\top \mathbf{1}_N)(\mathbf{T}^\top \mathbf{1}_N)^\top (\overline{\mathbf{C}}(\mathbf{T}) \odot \overline{\mathbf{C}}(\mathbf{T}))\right). \tag{47}$$

Consequently for $\mathbf{T} \in \mathcal{U}_n(\mathbf{h}_X)$ such that $\forall i \in [\![n]\!], \mathbf{T}_{:,i} \neq 0$ we have

$$F(\mathbf{T}, \overline{\mathbf{C}}(\mathbf{T})) = \mathrm{cte} - \sum_{ij} \frac{(\mathbf{T}_{:,i}^\top \mathbf{C}_X(\mathbf{X})\mathbf{T}_{:,j})^2}{(\mathbf{T}_{:,i}^\top \mathbf{1}_N)(\mathbf{T}_{:,j}^\top \mathbf{1}_N)} = \mathrm{cte} - \|\mathrm{diag}(\mathbf{T}^\top \mathbf{1}_N)^{-\frac{1}{2}} \mathbf{T}^\top \mathbf{C}_X(\mathbf{X})\mathbf{T}\,\mathrm{diag}(\mathbf{T}^\top \mathbf{1}_N)^{-\frac{1}{2}}\|_F^2. \tag{48}$$

It just remains to demonstrate the continuity of G. We consider for $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}_+^N \times \mathbb{R}_+^N$ the function

$$g(\mathbf{x}, \mathbf{y}) = \begin{cases} \frac{(\mathbf{x}^\top \mathbf{C}_X(\mathbf{X})\mathbf{y})^2}{\|\mathbf{x}\|_1 \|\mathbf{y}\|_1} & \text{when } \mathbf{x} \neq 0 \text{ and } \mathbf{y} \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{49}$$

and we show that g is continuous. For $(\mathbf{x}, \mathbf{y}) \neq (0, 0)$ this is clear. Now using that

$$0 \leq g(\mathbf{x}, \mathbf{y}) = \frac{(\sum_{ij}[\mathbf{C}_X(\mathbf{X})]_{ij} x_i y_j)^2}{(\sum_i x_i)(\sum_j y_j)} \leq \|\mathbf{C}_X(\mathbf{X})\|_\infty^2 \frac{(\sum_i x_i)^2 (\sum_j y_j)^2}{(\sum_i x_i)(\sum_j y_j)} = \|\mathbf{C}_X(\mathbf{X})\|_\infty^2 \|\mathbf{x}\|_1 \|\mathbf{y}\|_1, \tag{50}$$

this shows $\lim_{\mathbf{x} \to 0} g(\mathbf{x}, \mathbf{y}) = 0 = g(0, \mathbf{y})$ and $\lim_{\mathbf{y} \to 0} g(\mathbf{x}, \mathbf{y}) = 0 = g(\mathbf{x}, 0)$. Now for $\mathbf{T} \in \mathcal{U}_n(\mathbf{h}_X)$ we have $G(\mathbf{T}) = \mathrm{cte} - \sum_{ij} g(\mathbf{T}_{:,i}, \mathbf{T}_{:,j})$ which defines a continuous function.

$\square$

To prove the theorem we will first prove that the function $G : \mathbf{T} \to \min_{\overline{\mathbf{C}} \in \mathbb{R}^{n \times n}} E_L(\mathbf{C}_X(\mathbf{X}), \overline{\mathbf{C}}, \mathbf{T})$ is *concave* on $\mathcal{U}_n^+(\mathbf{h}_X)$ and by a continuity argument it will be concave on $\mathcal{U}_n(\mathbf{h}_X)$. The concavity will allow us to prove that the minimum of G is achieved in an extreme point of $\mathcal{U}_n(\mathbf{h}_X)$ which is a membership matrix.

Proposition C.2. *Let $\mathbf{h}_\mathbf{X} \in \Sigma_N, L = L_2$ and suppose that $\mathbf{C}_X(\mathbf{X})$ is CPD or CND. Then the function $G : \mathbf{T} \to \min_{\overline{\mathbf{C}} \in \mathbb{R}^{n \times n}} E_L(\mathbf{C}_X(\mathbf{X}), \overline{\mathbf{C}}, \mathbf{T})$ is concave on $\mathcal{U}_n(\mathbf{h}_X)$. Consequently Theorem 4.1 holds.*

Proof. We recall that $F(\mathbf{T}, \overline{\mathbf{C}}) := E_L(\mathbf{C}_X(\mathbf{X}), \overline{\mathbf{C}}, \mathbf{T})$ and $G(\mathbf{T}) = F(\mathbf{T}, \overline{\mathbf{C}}(\mathbf{T}))$. From Lemma C.1 we know that $\overline{\mathbf{C}}(\mathbf{T})$ is a minimizer of $\overline{\mathbf{C}} \to F(\mathbf{T}, \overline{\mathbf{C}})$ hence it satisfies the first order conditions $\partial_2 F(\mathbf{T}, \overline{\mathbf{C}}(\mathbf{T})) = 0$. Every quantity is differentiable on $\mathcal{U}_n^+(\mathbf{h}_X)$. Hence, taking the derivative of G and using the first order conditions

$$\nabla G(\mathbf{T}) = \partial_1 F(\mathbf{T}, \overline{\mathbf{C}}(\mathbf{T})) + \partial_2 F(\mathbf{T}, \overline{\mathbf{C}}(\mathbf{T}))[\nabla \overline{\mathbf{C}}(\mathbf{T})] = \partial_1 F(\mathbf{T}, \overline{\mathbf{C}}(\mathbf{T})). \tag{51}$$

We will found the expression of this gradient. In the proof of Lemma C.1 we have seen that

$$
\begin{aligned}
F(\mathbf{T}, \overline{\mathbf{C}}) &= \text{cte} + \text{Tr}((\mathbf{T}^\top \mathbf{1}_N)(\mathbf{T}^\top \mathbf{1}_N)^\top (\overline{\mathbf{C}} \odot \overline{\mathbf{C}})) - 2\,\text{Tr}(\mathbf{T}^\top \mathbf{C}_X(\mathbf{X})\mathbf{T}\overline{\mathbf{C}}) \\
&= \text{cte} + \text{Tr}(\mathbf{T}^\top \mathbf{1}_N \mathbf{1}_N^\top \mathbf{T}(\overline{\mathbf{C}} \odot \overline{\mathbf{C}})) - 2\,\text{Tr}(\mathbf{T}^\top \mathbf{C}_X(\mathbf{X})\mathbf{T}\overline{\mathbf{C}}) .
\end{aligned}
\tag{52}
$$

Using that the derivative of $\mathbf{T} \to \text{Tr}(\mathbf{T}^\top \mathbf{A}\mathbf{T}\mathbf{B})$ is $\mathbf{A}^\top \mathbf{T}\mathbf{B}^\top + \mathbf{A}\mathbf{T}\mathbf{B}$ and that $\mathbf{C}_X(\mathbf{X})$ is symmetric we get

$$
\partial_1 F(\mathbf{T}, \overline{\mathbf{C}}) = \mathbf{1}_N \mathbf{1}_N^\top \mathbf{T}(\overline{\mathbf{C}} \odot \overline{\mathbf{C}})^\top + \mathbf{1}_N \mathbf{1}_N^\top \mathbf{T}(\overline{\mathbf{C}} \odot \overline{\mathbf{C}}) - 2\mathbf{C}_X(\mathbf{X})\mathbf{T}\overline{\mathbf{C}}^\top - 2\mathbf{C}_X(\mathbf{X})\mathbf{T}\overline{\mathbf{C}} .
\tag{53}
$$

Finally, applying to the symmetric matrix $\overline{\mathbf{C}} = \overline{\mathbf{C}}(\mathbf{T})$

$$
\nabla G(\mathbf{T}) = \partial_1 F(\mathbf{T}, \overline{\mathbf{C}}(\mathbf{T})) = 2\left(\mathbf{1}_N \mathbf{1}_N^\top \mathbf{T}(\overline{\mathbf{C}}(\mathbf{T}) \odot \overline{\mathbf{C}}(\mathbf{T})) - 2\mathbf{C}_X(\mathbf{X})\mathbf{T}\overline{\mathbf{C}}(\mathbf{T})\right) .
\tag{54}
$$

In what follows we define

$$
D(\mathbf{T}) := \text{diag}(\mathbf{T}^\top \mathbf{1}_N)^{-1} \in \mathbb{R}^{n \times n} ,
\tag{55}
$$

when applicable. Using the expression of the gradient we will show that G is concave on $\mathcal{U}_n^+(\mathbf{h}_X)$ and we will conclude by a continuity argument on $\mathcal{U}_n(\mathbf{h}_X)$. Take $(\mathbf{P}, \mathbf{Q}) \in \mathcal{U}_n^+(\mathbf{h}_X) \times \mathcal{U}_n^+(\mathbf{h}_X)$ we will prove

$$
G(\mathbf{P}) - G(\mathbf{Q}) - \langle \nabla G(\mathbf{Q}), \mathbf{P} - \mathbf{Q} \rangle \leq 0 .
\tag{56}
$$

From Lemma C.1 we have the expression (since $\mathbf{P} \in \mathcal{U}_n^+(\mathbf{h}_X)$)

$$
G(\mathbf{P}) = \text{cte} - \|D(\mathbf{P})^{\frac{1}{2}} \mathbf{P}^\top \mathbf{C}_X(\mathbf{X})\mathbf{P}D(\mathbf{P})^{\frac{1}{2}}\|_F^2 ,
\tag{57}
$$

(same for $G(\mathbf{Q})$) and

$$
\overline{\mathbf{C}}(\mathbf{Q}) = D(\mathbf{Q})\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q}) .
\tag{58}
$$

We will now calculate $\langle \nabla G(\mathbf{Q}), \mathbf{P} \rangle$ which involves $\langle \mathbf{1}_N \mathbf{1}_N^\top \mathbf{Q}(\overline{\mathbf{C}}(\mathbf{Q}) \odot \overline{\mathbf{C}}(\mathbf{Q})), \mathbf{P} \rangle$ and $\langle \mathbf{C}_X(\mathbf{X})\mathbf{Q}\overline{\mathbf{C}}(\mathbf{Q}), \mathbf{P} \rangle$. For the first term we have

$$
\begin{aligned}
\langle \mathbf{1}_N \mathbf{1}_N^\top \mathbf{Q}(\overline{\mathbf{C}}(\mathbf{Q}) \odot \overline{\mathbf{C}}(\mathbf{Q})), \mathbf{P} \rangle &= \text{Tr}(\mathbf{P}^\top \mathbf{1}_N \mathbf{1}_N^\top \mathbf{Q}\overline{\mathbf{C}}(\mathbf{Q})^{\odot 2}) = \text{Tr}(\mathbf{1}_N^\top \mathbf{Q}\overline{\mathbf{C}}(\mathbf{Q})^{\odot 2}\mathbf{P}^\top \mathbf{1}_N) \\
&= (\mathbf{Q}^\top \mathbf{1}_N)^\top (\overline{\mathbf{C}}(\mathbf{Q}) \odot \overline{\mathbf{C}}(\mathbf{Q}))\mathbf{P}^\top \mathbf{1}_N \\
&= \text{Tr}(\overline{\mathbf{C}}(\mathbf{Q})\,\text{diag}(\mathbf{Q}^\top \mathbf{1}_N)\overline{\mathbf{C}}(\mathbf{Q})\,\text{diag}(\mathbf{P}^\top \mathbf{1}_N)) \\
&= \text{Tr}\left([D(\mathbf{Q})\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})]D(\mathbf{Q})^{-1}[D(\mathbf{Q})\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})]D(\mathbf{P})^{-1}\right) \\
&= \text{Tr}(D(\mathbf{Q})\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})D(\mathbf{P})^{-1}) \\
&= \text{Tr}(D(\mathbf{P})^{-\frac{1}{2}}[D(\mathbf{Q})\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})]D(\mathbf{P})^{-\frac{1}{2}}) \\
&= \text{Tr}(D(\mathbf{P})^{-\frac{1}{2}}D(\mathbf{Q})\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}}D(\mathbf{Q})^{\frac{1}{2}}\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})D(\mathbf{P})^{-\frac{1}{2}}) \\
&= \langle D(\mathbf{P})^{-\frac{1}{2}}D(\mathbf{Q})\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}}, D(\mathbf{P})^{-\frac{1}{2}}D(\mathbf{Q})\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}} \rangle \\
&= \|D(\mathbf{P})^{-\frac{1}{2}}D(\mathbf{Q})\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}}\|_F^2 .
\end{aligned}
\tag{59}
$$

For the second term

$$
\begin{aligned}
\langle \mathbf{C}_X(\mathbf{X})\mathbf{Q}\overline{\mathbf{C}}(\mathbf{Q}), \mathbf{P} \rangle &= \text{Tr}(\mathbf{P}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})) \\
&= \text{Tr}(D(\mathbf{Q})^{\frac{1}{2}}\mathbf{P}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}}D(\mathbf{Q})^{\frac{1}{2}}\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}}) \\
&= \langle D(\mathbf{Q})^{\frac{1}{2}}\mathbf{P}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}}, D(\mathbf{Q})^{\frac{1}{2}}\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}} \rangle \\
&= \langle D(\mathbf{Q})^{\frac{1}{2}}\mathbf{P}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}}, D(\mathbf{P})^{\frac{1}{2}}D(\mathbf{Q})^{-\frac{1}{2}}D(\mathbf{P})^{-\frac{1}{2}}D(\mathbf{Q})\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}} \rangle \\
&= \langle D(\mathbf{P})^{\frac{1}{2}}\mathbf{P}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}}, D(\mathbf{P})^{-\frac{1}{2}}D(\mathbf{Q})\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}} \rangle .
\end{aligned}
\tag{60}
$$

This gives

$$
\begin{aligned}
\langle \nabla G(\mathbf{Q}), \mathbf{P} \rangle &= 2\left(\|D(\mathbf{P})^{-\frac{1}{2}}D(\mathbf{Q})\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}}\|_F^2 - 2\langle D(\mathbf{P})^{\frac{1}{2}}\mathbf{P}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}}, D(\mathbf{P})^{-\frac{1}{2}}D(\mathbf{Q})\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}} \rangle\right) \\
&= 2\left(\|D(\mathbf{P})^{-\frac{1}{2}}D(\mathbf{Q})\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}} - D(\mathbf{P})^{\frac{1}{2}}\mathbf{P}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}}\|_F^2 - \|D(\mathbf{P})^{\frac{1}{2}}\mathbf{P}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}}\|_F^2\right) .
\end{aligned}
\tag{61}
$$

From this equation we get directly that

$$\langle \nabla G(\mathbf{Q}), \mathbf{Q} \rangle = -2\|D(\mathbf{Q})^{\frac{1}{2}}\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}}\|_F^2 \text{ and } \langle \nabla G(\mathbf{Q}), \mathbf{P} \rangle \geq -2\|D(\mathbf{P})^{\frac{1}{2}}\mathbf{P}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}}\|_F^2. \quad (62)$$

Hence

$$
\begin{aligned}
G(\mathbf{P}) - G(\mathbf{Q}) - \langle \nabla G(\mathbf{Q}), \mathbf{P} - \mathbf{Q} \rangle &= -\|D(\mathbf{P})^{\frac{1}{2}}\mathbf{P}^\top \mathbf{C}_X(\mathbf{X})\mathbf{P}D(\mathbf{P})^{\frac{1}{2}}\|_F^2 + \|D(\mathbf{Q})^{\frac{1}{2}}\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}}\|_F^2 \\
&\quad - \langle \nabla G(\mathbf{Q}), \mathbf{P} \rangle + \langle \nabla G(\mathbf{Q}), \mathbf{Q} \rangle \\
&\stackrel{eq.\ (62)}{=} -\|D(\mathbf{P})^{\frac{1}{2}}\mathbf{P}^\top \mathbf{C}_X(\mathbf{X})\mathbf{P}D(\mathbf{P})^{\frac{1}{2}}\|_F^2 - \|D(\mathbf{Q})^{\frac{1}{2}}\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}}\|_F^2 \\
&\quad - \langle \nabla G(\mathbf{Q}), \mathbf{P} \rangle \\
&\stackrel{eq.\ (62)}{\leq} -\|D(\mathbf{P})^{\frac{1}{2}}\mathbf{P}^\top \mathbf{C}_X(\mathbf{X})\mathbf{P}D(\mathbf{P})^{\frac{1}{2}}\|_F^2 - \|D(\mathbf{Q})^{\frac{1}{2}}\mathbf{Q}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}}\|_F^2 \\
&\quad + 2\|D(\mathbf{P})^{\frac{1}{2}}\mathbf{P}^\top \mathbf{C}_X(\mathbf{X})\mathbf{Q}D(\mathbf{Q})^{\frac{1}{2}}\|_F^2.
\end{aligned}
\tag{63}
$$

We note $\mathbf{U} = \mathbf{P}D(\mathbf{P})\mathbf{P}^\top \in \mathbb{R}^{N \times N}$, $\mathbf{V} = \mathbf{Q}D(\mathbf{Q})\mathbf{Q}^\top \in \mathbb{R}^{N \times N}$, the previous calculus shows that

$$
\begin{aligned}
G(\mathbf{P}) - G(\mathbf{Q}) - \langle \nabla G(\mathbf{Q}), \mathbf{P} - \mathbf{Q} \rangle &\leq -\operatorname{Tr}(\mathbf{U}\mathbf{C}_X(\mathbf{X})\mathbf{U}\mathbf{C}_X(\mathbf{X})) - \operatorname{Tr}(\mathbf{V}\mathbf{C}_X(\mathbf{X})\mathbf{V}\mathbf{C}_X(\mathbf{X})) \\
&\quad + 2\operatorname{Tr}(\mathbf{V}\mathbf{C}_X(\mathbf{X})\mathbf{U}\mathbf{C}_X(\mathbf{X})),
\end{aligned}
\tag{64}
$$

Now note that

$$\mathbf{U}^\top \mathbf{1}_N = \mathbf{P}D(\mathbf{P})\mathbf{P}^\top \mathbf{1}_N = \mathbf{P}\operatorname{diag}(\mathbf{P}^\top \mathbf{1}_N)^{-1}\mathbf{P}^\top \mathbf{1}_N = \mathbf{P}\mathbf{1}_N = \mathbf{h}_X. \quad (65)$$

Since $\mathbf{U}$ is symmetric we also have $\mathbf{U}\mathbf{1}_N = \mathbf{h}_X$ and similarly we have the same result for $\mathbf{V}$. Overall $\mathbf{V}^\top \mathbf{1}_N = \mathbf{U}^\top \mathbf{1}_N$ and $\mathbf{V}\mathbf{1}_N = \mathbf{U}\mathbf{1}_N$. Since $\mathbf{C}_X(\mathbf{X})$ is CPD or CND we can apply Lemma C.3 below which proves that $-\operatorname{Tr}(\mathbf{U}\mathbf{C}_X(\mathbf{X})\mathbf{U}\mathbf{C}_X(\mathbf{X})) - \operatorname{Tr}(\mathbf{V}\mathbf{C}_X(\mathbf{X})\mathbf{V}\mathbf{C}_X(\mathbf{X})) + 2\operatorname{Tr}(\mathbf{V}\mathbf{C}_X(\mathbf{X})\mathbf{U}\mathbf{C}_X(\mathbf{X})) \leq 0$ and consequently that G is concave on $\mathcal{U}_n^+(\mathbf{h}_X)$. We now use the continuity of G to prove that it is concave on $\mathcal{U}_n(\mathbf{h}_X)$.

Take $\mathbf{P} \in \mathcal{U}_n^+(\mathbf{h}_X)$ and $\mathbf{Q} \in \mathcal{U}_n(\mathbf{h}_X) \setminus \mathcal{U}_n^+(\mathbf{h}_X)$ i.e. there exists $k \in [\![n]\!]$ such that $\mathbf{Q}_{:,k} = 0$. Without loss of generality we suppose $k = 1$. Consider for $m \in \mathbb{N}^*$ the matrix $\mathbf{Q}^{(m)} = (\frac{1}{m}\mathbf{1}_N, \mathbf{Q}_{:,2}, \cdots, \mathbf{Q}_{:,n})$. Then $\mathbf{Q}^{(m)} \to \mathbf{Q}$ as $m \to +\infty$. Also since $\mathbf{Q}^{(m)} \in \mathcal{U}_n^+(\mathbf{h}_X)$ we have by concavity of G

$$G((1-\lambda)\mathbf{P} + \lambda\mathbf{Q}^{(m)}) \geq (1-\lambda)G(\mathbf{P}) + \lambda G(\mathbf{Q}^{(m)}), \quad (66)$$

for any $\lambda \in [0, 1]$. Taking the limit as $m \to \infty$ gives, by continuity of G,

$$G((1-\lambda)\mathbf{P} + \lambda\mathbf{Q}) \geq (1-\lambda)G(\mathbf{P}) + \lambda G(\mathbf{Q}), \quad (67)$$

and hence G is concave on $\mathcal{U}_n(\mathbf{h}_X)$. This proves Theorem 4.1. Indeed the minimization of $\mathbf{T} \in \mathcal{U}_n(\mathbf{h}_X) \to \min_{\overline{\mathbf{C}}} E_L(\mathbf{C}_X(\mathbf{X}), \overline{\mathbf{C}}, \mathbf{T})$ is a minimization of a concave function over a polytope, hence admits an extremity of $\mathcal{U}_n(\mathbf{h}_X)$ as minimizer. But these extremities are membership matrices as they can be described as $\{\operatorname{diag}(\mathbf{h}_X)\mathbf{P} : \mathbf{P} \in \{0,1\}^{N \times n}, \mathbf{P}^\top \mathbf{1}_n = \mathbf{1}_N\}$ (Cao et al., 2022). $\qquad\square$

Lemma C.3. *Let $\mathbf{C} \in \mathbb{R}^{N \times N}$ be a CPD or CND matrix. Then for any $(\mathbf{P}, \mathbf{Q}) \in \mathbb{R}^{N \times N} \times \mathbb{R}^{N \times N}$ such that $\mathbf{P}^\top \mathbf{1}_N = \mathbf{Q}^\top \mathbf{1}_N$ and $\mathbf{P}\mathbf{1}_N = \mathbf{Q}\mathbf{1}_N$ we have*

$$\operatorname{Tr}(\mathbf{P}^\top \mathbf{C}\mathbf{Q}\mathbf{C}) \leq \frac{1}{2}(\operatorname{Tr}(\mathbf{P}^\top \mathbf{C}\mathbf{P}\mathbf{C}) + \operatorname{Tr}(\mathbf{Q}^\top \mathbf{C}\mathbf{Q}\mathbf{C})). \quad (68)$$

Proof. First, since $\mathbf{C}$ is symmetric,

$$
\begin{aligned}
\operatorname{Tr}\left((\mathbf{P} - \mathbf{Q})^\top \mathbf{C}(\mathbf{P} - \mathbf{Q})\mathbf{C}\right) &= \operatorname{Tr}(\mathbf{P}^\top \mathbf{C}\mathbf{P}\mathbf{C} - \mathbf{P}^\top \mathbf{C}\mathbf{Q}\mathbf{C} - \mathbf{Q}^\top \mathbf{C}\mathbf{P}\mathbf{C} + \mathbf{Q}^\top \mathbf{C}\mathbf{Q}\mathbf{C}) \\
&= \operatorname{Tr}(\mathbf{P}^\top \mathbf{C}\mathbf{P}\mathbf{C}) + \operatorname{Tr}(\mathbf{Q}^\top \mathbf{C}\mathbf{Q}\mathbf{C}) - 2\operatorname{Tr}(\mathbf{P}^\top \mathbf{C}\mathbf{Q}\mathbf{C}).
\end{aligned}
\tag{69}
$$

We note $\mathbf{U} = \mathbf{P} - \mathbf{Q}$. Since $\mathbf{P}^\top \mathbf{1}_N = \mathbf{Q}^\top \mathbf{1}_N$ we have $\mathbf{U}^\top \mathbf{1}_N = 0$. In the same way $\mathbf{U}\mathbf{1}_N = 0$. We introduce $\mathbf{H} = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top$ the centering matrix. Note that

$$\mathbf{H}\mathbf{U}\mathbf{H} = (\mathbf{U} - \frac{1}{N}\mathbf{1}_N(\mathbf{1}_N^\top \mathbf{U}))\mathbf{H} = \mathbf{U}\mathbf{H} = \mathbf{U} - \frac{1}{N}(\mathbf{U}\mathbf{1}_N)\mathbf{1}_N^\top = \mathbf{U}. \quad (70)$$

Also $\mathbf{C}$ is CPD if and only if $\mathbf{HCH}$ is positive semi-definite (PSD). Indeed if $\mathbf{HCH}$ is PSD then take $\mathbf{x}$ such that $\mathbf{x}^\top \mathbf{1}_N = 0$. We then have $\mathbf{Hx} = \mathbf{x}$ and thus $\mathbf{x}^\top \mathbf{Cx} = \mathbf{x}^\top(\mathbf{HCH})\mathbf{x} \geq 0$. On the other hand when $\mathbf{C}$ is CPD then take any $\mathbf{x}$ and see that $\mathbf{x}^\top \mathbf{HCHx} = (\mathbf{Hx})^\top \mathbf{C}(\mathbf{Hx})$. But $(\mathbf{Hx})^\top \mathbf{1}_N = \mathbf{x}^\top(\mathbf{H}^\top \mathbf{1}_N) = 0$. So $(\mathbf{Hx})^\top \mathbf{C}(\mathbf{Hx}) \geq 0$.

By hypothesis $\mathbf{C}$ is CPD so $\mathbf{HCH}$ is PSD and symmetric, so it has a square root. But using eq. (70) we get

$$
\begin{aligned}
\mathrm{Tr}\left((\mathbf{P} - \mathbf{Q})^\top \mathbf{C}(\mathbf{P} - \mathbf{Q})\mathbf{C}\right) = \mathrm{Tr}(\mathbf{U}^\top \mathbf{CUC}) &= \mathrm{Tr}(\mathbf{HU}^\top \mathbf{HCHUHC}) \\
&= \mathrm{Tr}(\mathbf{U}^\top(\mathbf{HCH})\mathbf{U}(\mathbf{HCH})) \\
&= \|(\mathbf{HCH})^{\frac{1}{2}}\mathbf{U}(\mathbf{HCH})^{\frac{1}{2}}\|_F^2 \geq 0,
\end{aligned}
\tag{71}
$$

For the CND case is suffices to use that $\mathbf{C}$ is CND if and only if $-\mathbf{C}$ is CPD and that $\mathrm{Tr}\left((\mathbf{P} - \mathbf{Q})^\top \mathbf{C}(\mathbf{P} - \mathbf{Q})\mathbf{C}\right) = \mathrm{Tr}\left((\mathbf{P} - \mathbf{Q})^\top(-\mathbf{C})(\mathbf{P} - \mathbf{Q})(-\mathbf{C})\right)$ which concludes the proof. $\qquad\square$

D. Algorithmic details

We detail in the following the algorithms mentioned in Section 4 to address the semi-relaxed GW divergence computation in our Block Coordinate Descent algorithm for the DistR problem. We begin with details on the computation of an equivalent objective function and its gradient, potentially under low-rank assumptions over structures $\mathbf{C}$ and $\overline{\mathbf{C}}$.

D.1. Objective function and gradient computation.

Problem statement. Let consider any matrices $\mathbf{C} \in \mathbb{R}^{n \times n}$, $\overline{\mathbf{C}} \in \mathbb{R}^{m \times m}$, and a probability vector $\mathbf{h} \in \Sigma_n$. In all our use cases, we considered inner losses $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ which can be decomposed following Proposition 1 in (Peyré et al., 2016). Namely we assume the existence of functions f_1, f_2, h_1, h_2 such that

$$
\forall a, b \in \Omega^2, \quad L(a, b) = f_1(a) + f_2(b) - h_1(a)h_2(b) \tag{72}
$$

More specifically we considered

$$
\begin{aligned}
L_2(a, b) = (a - b)^2 &\implies f_1(a) = a^2, \quad f_2(b) = b^2, \quad h_1(a) = a, \quad h_2(b) = 2b \\
L_{KL}(a, b) = a \log \frac{a}{b} - a + b &\implies f_1(a) = a \log a - a, \quad f_2(b) = b, \quad h_1(a) = a, \quad h_2(b) = \log b
\end{aligned}
\tag{L2}
$$

In this setting, we proposed to solve for the equivalent problem to srGW_L :

$$
\min_{\mathbf{T} \in \mathcal{U}_n(\mathbf{h})} F(\mathbf{T}) \tag{srGW-2}
$$

where the objective function reads as

$$
\begin{aligned}
F(\mathbf{T}) &:= \langle F_1(\overline{\mathbf{C}}, \mathbf{T}) - F_2(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{T}), \mathbf{T} \rangle \\
&= \langle \mathbf{1}_N \mathbf{1}_n^\top \mathbf{T}^\top f_2(\overline{\mathbf{C}}^\top), \mathbf{T} \rangle - \langle h_1(\mathbf{C})\mathbf{T}h_2(\overline{\mathbf{C}})^\top, \mathbf{T} \rangle
\end{aligned}
\tag{73}
$$

Problem srGW-2 is usually a non-convex QP with Hessian $\mathcal{H} = f_2(\overline{\mathbf{C}}) \otimes \mathbf{1}\mathbf{1}^\top - h_2(\overline{\mathbf{C}}) \otimes_K h_1(\mathbf{C})$. In all cases this equivalent form is interesting as it avoids computing the constant term $\langle f_1(\mathbf{C}), \mathbf{hh}^\top \rangle$ that requires $O(n^2)$ operations in all cases.

The gradient of F w.r.t $\mathbf{T}$ then reads as

$$
\nabla_{\mathbf{T}} F(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{T}) = F_1(\overline{\mathbf{C}}, \mathbf{T}) + F_1(\overline{\mathbf{C}}^\top, \mathbf{T}) - F_2(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{T}) - F_2(\mathbf{C}^\top, \overline{\mathbf{C}}^\top, \mathbf{T}) \tag{74}
$$

When $C_X(\mathbf{X})$ and $C_Z(\mathbf{Z})$ are symmetric, which is the case in all our experiments, this gradient reduces to $\nabla_{\mathbf{T}} F = 2(F_1 - F_2)$.

Low-rank factorization. Inspired from the work of Scetbon et al. (2022), we propose implementations of srGW that can leverage the low-rank nature of $C_X(\mathbf{X})$ and $C_Z(\mathbf{Z})$. Let us assume that both structures can be exactly decomposed as

follows, $\mathbf{C}_X(\mathbf{X}) = \mathbf{A}_1\mathbf{A}_1^\top$ where $\mathbf{A} \in \mathbb{R}^{N \times r}$ and $\mathbf{C}_Z(\mathbf{Z}) = \mathbf{B}_1\mathbf{B}_1^\top$ with $\mathbf{B} \in \mathbb{R}^{n \times s}$, such that $r << N$ and $s << n$. For both inner losses L we make use of the following factorization:

$\underline{L = L_2}$: Computing the first term F_1 coming for the optimized second marginal can not be factored because $f_2(\mathbf{C}_Z(\mathbf{Z})) = \mathbf{C}_Z(\mathbf{Z})^2$, does not preserve the low-rank structure. Hence computing $F_2(\cdot, \cdot, \mathbf{T})$, following this paranthesis $\mathbf{1}_N((\mathbf{T}\mathbf{1}_n)^\top)f_2(\overline{\mathbf{C}})^\top)$ in $O(Nn + n^2)$ operation. And its scalar product with $\mathbf{T}$ comes down to $O(Nn)$ operations. Then Computing $F_2(\mathbf{T})$ and its scalar product with $\mathbf{T}$ can be done following the development of Scetbon et al. (2022, Section 3) for the corresponding GW problem, in $O(Nn(r + s) + rs(N + n))$ operations. So the overall complexity at is $O(Nn(r + s) + rs(N + n) + n^2)$.

$\underline{L = L_{KL}}$: In this setting $f_2(\overline{\mathbf{C}}) = \overline{\mathbf{C}}$ and $h_1(\mathbf{C}) = \mathbf{C}$ naturally preserves the low-rank nature of input matrices, but $h_2(\overline{\mathbf{C}}) = \log(\overline{\mathbf{C}})$ does not. So computing the first term F_1, can be performed following this paranthesis order $\mathbf{1}_N((\mathbf{1}_n^\top(\mathbf{T}^\top\mathbf{A}))\mathbf{A}^\top)$ in $O(Nnr)$ operations. While the second term F_2 should be computed following this order $\mathbf{A}((\mathbf{A}^\top\mathbf{T})\log(\overline{\mathbf{C}}))$ in $O(Nnr + rn^2)$ operations. While their respective scalar product can be computed in $O(Nn)$. So the overall complexity is $O(Nnr + n^2r)$.

D.2. Solvers.

We develop next our extension of both the Mirror Descent and Conditional Gradient solvers first introduced in Vincent-Cuaz et al. (2022a), for any inner loss L that decomposes as in equation 72 .

Mirror Descent algorithm.   This solver comes down to solve for the *exact* srGW problem using mirror-descent scheme w.r.t the KL geometry. At each iteration (i), the solver comes down to, first computing the gradient $\nabla_\mathbf{T}F(\mathbf{T}^{(i)})$ given in equation 74 evaluated in $\mathbf{T}^{(i)}$, then updating the transport plan using the following closed-form solution to a KL projection:

$$\mathbf{T}^{(i+1)} \leftarrow \text{diag}\left(\frac{\mathbf{h}}{\mathbf{K}^{(i)}\mathbf{1}_n}\right)\mathbf{K}^{(i)} \tag{75}$$

where $\mathbf{K}^{(i)} = \exp\left(\nabla_\mathbf{T}F(\mathbf{T}^{(i)}) - \varepsilon \log(\mathbf{T}^{(i)})\right)$ and $\varepsilon > 0$ is an hyperparameter to tune. Proposition 3 and Lemma 7 in Vincent-Cuaz (2023, Chapter 6) provides that the Mirror-Descent algorithm converges to a stationary point non-asymptotically when $L = L_2$. A quick inspection of the proof suffices to see that this convergence holds for any losses L satisfying equation 72, up to adaptation of constants involved in the Lemma.

Conditional Gradient algorithm.   This algorithm, known to converge to local optimum (Lacoste-Julien, 2016), iterates over the 3 steps summarized in Algorithm 1: .

Algorithm 1 CG solver for srGW$_L$

1: **repeat**
2:   $\mathbf{F}^{(i)} \leftarrow$ Compute gradient w.r.t $\mathbf{T}$ of equation 74.
3:   $\mathbf{X}^{(i)} \leftarrow \min_{\substack{\mathbf{X}\mathbf{1}_m = \mathbf{h} \\ \mathbf{X} \geq 0}} \langle \mathbf{X}, \mathbf{F}^{(i)}\rangle$
4:   $\mathbf{T}^{(i+1)} \leftarrow (1 - \gamma^\star)\mathbf{T}^{(i)} + \gamma^\star\mathbf{X}^{(i)}$ with $\gamma^\star \in [0, 1]$ from exact-line search.
5: **until** convergence.

This algorithm consists in solving at each iteration (i) a linearization $\langle \mathbf{X}, \mathbf{F}^{(i)}\rangle$ of the problem equation srGW-2 where $\mathbf{F}(\mathbf{T}^{(i)})$ is the gradient of the objective in equation 74. The solution of the linearized problem provides a *descent direction* $\mathbf{X}^{(i)} - \mathbf{T}^{(i)}$, and a line-search whose optimal step can be found in closed form to update the current solution $\mathbf{T}^{(i)}$. We detail in the following this line-search step for any loss that can be decomposed as in equation 72. It comes down for any $\mathbf{T} \in \mathcal{U}_n(\mathbf{h})$, to solve the following problem:

$$\gamma = \underset{\gamma \in [0,1]}{\arg\min}\, g(\gamma) := F(\mathbf{T} + \gamma(\mathbf{X} - \mathbf{T})) \tag{76}$$

Observe that this objective function can be developed as a second order polynom $g(\gamma) = a\gamma^2 + b\gamma + c$. To find an optimal γ it suffices to express coefficients a and b to conclude using Algorithm 2 in Vayer et al. (2018).

Denoting $\mathbf{X}^\top \mathbf{1}_n = \mathbf{q}_X$ and $\mathbf{T}^\top \mathbf{1}_n = \mathbf{q}_T$ and following equation 73, we have

$$
\begin{aligned}
a &= \langle \mathbf{1}_n(\mathbf{q}_X - \mathbf{q}_T)^\top f_2(\overline{\mathbf{C}})^\top - h_1(\mathbf{C})(\mathbf{X} - \mathbf{T})h_2(\overline{\mathbf{C}})^\top, \mathbf{X} - \mathbf{T} \rangle \\
&= \langle F_1(\mathbf{X}) - F_1(\mathbf{T}) - F_2(\mathbf{X}) + \mathbf{F}(\mathbf{T}), \mathbf{X} - \mathbf{T} \rangle
\end{aligned}
\tag{77}
$$

Finally the coefficient b of the linear term is

$$
b = \langle F_1(\mathbf{T}) - F_2(\mathbf{T}), \mathbf{X} - \mathbf{T} \rangle + \langle F_1(\mathbf{X} - \mathbf{T}) - F_2(\mathbf{X} - \mathbf{T}), \mathbf{T} \rangle
\tag{78}
$$

E. Additional Details for Experiments

Datasets. We provide details about the datasets used in section 5. MNIST and Fashion-MNIST are taken from Torchvision (Marcel & Rodriguez, 2010), the COIL dataset can be downloaded at `https://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php`, the two SNAREseq datasets can be found in `https://github.com/rsinghlab/SCOT`, the Zeisel dataset at the following repository `https://github.com/solevillar/scGeneFit-python` and PBMC can be downloaded from Scanpy (Wolf et al., 2018). When the dimensionality of a dataset exceeds 50, we pre-process it by applying a PCA in dimension 50, as done in practice (Van der Maaten & Hinton, 2008).

Table 2: Dataset Sizes.

	Number of samples	Dimensionality	Number of classes
MNIST	10000	784	10
F-MNIST	10000	784	10
COIL	1440	16384	20
SNAREseq (chromatin)	1047	19	5
SNAREseq (gene expression)	1047	10	5
Zeisel	3005	5000	(8, 49)
PBMC	2638	1838	8

Weighted silhouette score. The silhouette score measures whether positions of points are in adequation with their predicted clusters. In the classical DR framework (see Section 2.1), the measure associated to embeddings $\mathbf{Z} \in \mathbb{R}^{N \times d}$ contains as many locations than the input measure, and their relative importance is considered uniform. Assume that $\mathbf{z}_i$ belongs to cluster $y_i \in \{1, ..., K\}$ acting as K classes. The classical silhouette score is then defined as $\mathcal{S}(\mathbf{Z}, \mathbf{Y}) = \frac{1}{N} \sum_{i \in [\![N]\!]} s_i(\mathbf{Z}, \mathbf{Y})$, *i.e* the (uniform) mean of the silhouette coefficient associate to each prototype i, based on two main components:

i) The mean intra-cluster distance a_i: *i.e* the mean of distances from $\mathbf{z}_i$ to other locations $\mathbf{z}_j$ that belongs to the same cluster $y_i = y_j$,

$$
a_i(\mathbf{Z}, \mathbf{Y}) = \frac{\sum_{j \in [\![N]\!]} \delta_{y_i = y_j} d(\mathbf{z}_j, \mathbf{z}_j)}{\sum_{j \in [\![N]\!]} \delta_{y_i = y_j}}
\tag{79}
$$

ii) The mean nearest-cluster distance b_i: *i.e* the mean distace between the point $\mathbf{z}_i$ and the nearest cluster $\mathbf{z}_i$ does not belong to. Assume that the nearest cluster is k with samples $\{\mathbf{z}_j | y_j = k\}$, then we have

$$
b_i(\mathbf{Z}, \mathbf{Y}) = \frac{\sum_{j \in [\![N]\!]} \delta_{y_j = k} d(\mathbf{z}_j, \mathbf{z}_i)}{\sum_{j \in [\![N]\!]} \delta_{y_j = k}}
\tag{80}
$$

Then the silhouette coefficient of sample $\mathbf{z}_i$ reads as $s_i(\mathbf{Z}, \mathbf{Y}) = \frac{b_i - a_i}{\max(a_i, b_i)}$.

In our case, the prototypes $\mathbf{z}_i$ are associated with a relative importance $w_i = [\mathbf{h}_Z]_i$. Hence to incorporate it within the classical silhouette score it suffices to interpret the aforementioned components as if distances were weighted with uniform weights $w_i = 1/N$. It comes to define the *weighted* mean intra-cluster and nearest-cluster distances that read

$$
\widetilde{a}_i(\mathbf{Z}, \mathbf{Y}, \mathbf{w}) = \frac{\sum_{j \in [\![N]\!]} \delta_{y_i = y_j} w_j d(\mathbf{z}_j, \mathbf{z}_j)}{\sum_{j \in [\![N]\!]} \delta_{y_i = y_j} w_j} \quad \text{and} \quad \widetilde{b}_i(\mathbf{Z}, \mathbf{Y}, \mathbf{w}) = \frac{\sum_{j \in [\![N]\!]} \delta_{y_j = k} w_j d(\mathbf{z}_j, \mathbf{z}_i)}{\sum_{j \in [\![N]\!]} \delta_{y_j = k} w_j}
\tag{81}
$$

such that the weighted silhouette coefficient is $\widetilde{s}_i(\mathbf{Z}, \mathbf{Y}, \mathbf{w}) = \frac{\widetilde{b}_i - \widetilde{a}_i}{\max(\widetilde{b}_i, \widetilde{a}_i)}$ and the weighted silhouette score reads as

$$\widetilde{S}(\mathbf{Z}, \mathbf{Y}, \mathbf{w}) = \sum_{i \in [\![N]\!]} w_i \widetilde{s}_i(\mathbf{Z}, \mathbf{Y}, \mathbf{w}) \tag{82}$$

Table 3: Silhouette scores $\mathcal{S}$ for configurations maximizing our joint DR/clustering Score $\overline{\mathcal{SH}}$.

	methods	$\mathbf{C}_X$ / $\mathbf{C}_Z$	MNIST	FMNIST	COIL	SNA1	SNA2	ZEI1	ZEI2	PBMC
$d=10$	DistR	$\langle,\rangle_{\mathbb{R}^p}$ / $\langle,\rangle_{\mathbb{R}^d}$	13.7 (0.3)	14.9 (0.8)	12.2 (1.9)	2.3 (0.0)	46.5 (11.6)	6.0 (4.0)	-23.1 (6.4)	53.0 (0.6)
	DistR$_\varepsilon$	-	13.6 (0.0)	11.0 (0.0)	14.1 (0.1)	14.4 (0.7)	15.4 (5.3)	19.1 (0.6)	-11.4 (0.6)	59.2 (0.1)
	DR→C	-	-1.0 (0.0)	-0.7 (0.0)	-1.2 (5.5)	14.2 (1.2)	-2.9 (9.7)	-5.6 (3.3)	-21.0 (5.3)	27.3 (4.2)
	C→DR	-	13.4 (0.0)	10.5 (0.0)	-18.2 (0.0)	12.2 (0.7)	56.8 (1.1)	18.3 (0.4)	-12.1 (0.5)	57.3 (0.0)
	COOT	NA	18.6 (9.3)	12.8 (23.8)	5.0 (3.2)	27.7 (15.8)	69.8 (8.5)	23.1 (6.8)	19.9 (6.3)	6.3 (7.7)
$d=2$	DistR	SEA / St.	21.4 (3.1)	19.3 (1.5)	37.2 (2.4)	27.9 (6.2)	54.0 (0.1)	23.8 (2.5)	-5.6 (2.9)	32.4 (3.5)
	DistR$_\varepsilon$	-	21.0 (0.9)	17.8 (2.6)	35.8 (2.1)	27.9 (6.2)	67.6 (0.2)	19.5 (0.7)	-22.4 (2.3)	51.5 (1.0)
	DR→C	-	16.9 (3.1)	19.2 (2.1)	23.7 (0.4)	15.8 (4.0)	56.9 (0.5)	15.7 (2.5)	-24.2 (2.1)	33.6 (2.6)
	C→DR	-	14.1 (2.2)	15.9 (1.1)	29.5 (1.1)	21.3 (8.0)	59.3 (0.2)	15.2 (2.7)	-31.9 (3.7)	60.3 (1.8)
	COOT	NA	0.5 (21.6)	-8.3 (3.0)	-15.6 (7.8)	30.4 (20.5)	27.6 (6.8)	12.6 (2.9)	1.0 (2.3)	-12.4 (10.3)
$d=2$	DistR	SEA / H-St.	0.0 (0.0)	1.1 (2.2)	-3.9 (4.6)	1.7 (3.3)	55.7 (7.3)	9.6 (12.3)	-50.3 (5.0)	8.4 (6.41)
	DistR$_\varepsilon$	-	-22.9 (1.9)	-18.9 (1.2)	-8.7 (3.5)	21.0 (20.2)	23.4 (3.9)	-5.7 (7.3)	-54.9 (2.2)	-0.8 (9.1)
	DR→C	-	-13.5 (4.2)	-9.0 (7.3)	-42.3 (2.4)	9.2 (16.9)	-7.7 (9.4)	22.5 (20.6)	-28.1 (9.4)	41.1 (21.4)
	C→DR	-	-27.3 (7.4)	-18.5 (2.1)	-2.6 (4.4)	23.1 (15.4)	0.0(0.0)	-15.0 (3.6)	-15.9 (4.9)	-23.7 (1.6)

Table 4: Homogeneity scores $\mathcal{H}$ for configurations maximizing our joint DR/clustering Score $\overline{\mathcal{SH}}$.

	methods	$\mathbf{C}_X$ / $\mathbf{C}_Z$	MNIST	FMNIST	COIL	SNA1	SNA2	ZEI1	ZEI2	PBMC
$d=10$	DistR	$\langle,\rangle_{\mathbb{R}^p}$ / $\langle,\rangle_{\mathbb{R}^d}$	97.0 (0.9)	90.6 (0.9)	98.6 (0.4)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	59.8 (3.2)	96.1 (1.3)
	DistR$_\varepsilon$	-	96.3 (0.0)	93.4 (0.0)	100.0 (0.0)	80.5 (0.0)	100.0 (0.0)	100.0 (0.0)	61.1 (0.0)	90.6 (0.0)
	DR→C	-	98.3 (3.4)	78.2 (0.0)	91.8 (5.2)	97.4 (5.1)	53.4 (7.7)	100.0 (0.0)	14.2 (15.1)	90.1 (0.2)
	C→DR	-	96.3 (0.0)	93.4 (0.0)	83.7 (0.0)	80.5 (0.0)	93.6 (0.0)	100.0 (0.0)	61.1 (0.0)	93.3 (0.0)
	COOT	NA	6.3 (2.0)	0.0 (0.0)	43.4 (1.0)	34.7 (9.8)	68.3 (10.0)	100.0 (0.0)	0.0 (0.0)	15.8 (3.9)
$d=2$	DistR	SEA / St.	94.3 (1.4)	93.9 (0.7)	98.0 (0.4)	96.4 (1.6)	100.0 (0.0)	100.0 (0.0)	48.0 (1.8)	98.8 (1.0)
	DistR$_\varepsilon$	-	95.3 (0.0)	92.3 (0.0)	100.0 (0.0)	96.4 (1.6)	97.6 (0.0)	100.0 (0.0)	57.5 (0.0)	97.0 (0.0)
	DR→C	-	91.2 (0.9)	91.8 (0.5)	98.2 (0.5)	95.7 (0.7)	98.1 (0.5)	100.0 (0.0)	51.9 (2.3)	94.7 (1.0)
	C→DR	-	95.3 (0.0)	92.0 (0.0)	98.5 (0.0)	95.0 (0.0)	100.0 (0.0)	100.0 (0.0)	57.5 (0.0)	96.6 (0.0)
	COOT	NA	2.0 (1.0)	3.9 (1.5)	42.8 (1.5)	0.0 (0.0)	49.7 (8.6)	100.0 (0.0)	0.0 (0.0)	14.1 (2.2)
$d=2$	DistR	SEA / H-St.	100.0 (0.0)	100.0 (0.0)	92.4 (1.7)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	58.4 (4.6)	92.2 (3.4)
	DistR$_\varepsilon$	-	95.3 (0.0)	92.2 (0.0)	93.5 (0.0)	100.0 (0.0)	95.5 (0.0)	100.0 (0.0)	54.3 (0.0)	96.3 (0.0)
	DRC	-	61.5 (15.3)	73.9 (17.8)	65.9 (2.2)	49.0 (12.1)	70.0(24.5)	100.0 (0.0)	47.1 (3.7)	64.2 (4.5)
	CDR	-	97.8 (0.0)	91.9 (0.0)	93.5 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	97.0 (0.0)
	CDR*	-	95.3 (0.0)	92.6 (0.0)	93.5 (0.0)	100.0 (0.0)	93.5 (0.0)	100.0 (0.0)	56.7 (0.0)	97.0 (0.0)