Interpolating between Clustering and Dimensionality Reduction with Gromov-Wasserstein

Hugues Van Assel[*]
ENS de Lyon, CNRS
UMPA UMR 5669
hugues.van_assel@ens-lyon.fr

Cédric Vincent-Cuaz[*]
EPFL, Lausanne
LTS4
cedric.vincent-cuaz@epfl.ch

Titouan Vayer
Univ. Lyon, ENS de Lyon, UCBL, CNRS, Inria
LIP UMR 5668
titouan.vayer@inria.fr

Rémi Flamary
École polytechnique, IP Paris, CNRS
CMAP UMR 7641
remi.flamary@polytechnique.edu

Nicolas Courty
Université Bretagne Sud, CNRS
IRISA UMR 6074
nicolas.courty@irisa.fr

Abstract

We present a versatile adaptation of existing dimensionality reduction (DR) objectives, enabling the simultaneous reduction of both sample and feature sizes. Correspondances between input and embedding samples are computed through a semi-relaxed Gromov-Wasserstein optimal transport (OT) problem. When the embedding sample size matches that of the input, our model recovers classical popular DR models. When the embedding's dimensionality is unconstrained, we show that the OT plan delivers a competitive hard clustering. We emphasize the importance of intermediate stages that blend DR and clustering for summarizing real data and apply our method to visualize datasets of images.

1 Introduction

Summarizing the information carried by a dataset in an unsupervised way is of utmost importance in modern machine learning pipelines [14]. Smaller representations of data offer numerous advantages, including improved pattern and structure recognition, as well as faster processing for downstream tasks [24, 5, 28]. To construct such representations, one can either reduce the sample size by aggregating points together (referred to as *clustering*) or reduce the feature dimensionality *i.e.* performing *dimensionality reduction* (DR). While both tasks are actively studied topics, very few works have proposed a consistent model to simultaneously perform clustering and DR.

Contributions. In this work, we provide a new framework for joint clustering and DR. The goal is to obtain a reduced representation in *both samples and features i.e.* a transformation $\boldsymbol{X} \in \mathbb{R}^{N \times p}$ to $\boldsymbol{Z} \in \mathbb{R}^{n \times d}$ where $n < N$ (clustering) and $d < p$ (DR). Doing so, we ensure that the low-dimensional embeddings align well with the class labels determined during clustering. In Section 2, we frame classical DR methods as minimizing a discrepancy between two aligned affinity matrices: $\boldsymbol{C}_X$ defining the dependencies among high-dimensional samples and $\boldsymbol{C}_Z$ focusing on low-dimensional

[*]Equal contribution.

Preprint. Under review.

ones. We then propose to augment this general objective using the Gromov-Wasserstein (GW) framework to enable matching affinities of different dimensions. When $\boldsymbol{C}_Z$ has fewer nodes than $\boldsymbol{C}_X$, computing a GW transport plan naturally amounts to a clustering of the input samples, aggregating them into *prototypes*. Therefore this model leads to a principled objective for simultaneously learning low-dimensional prototypes and their assignments to input samples. We show in Theorem 1 that, in the context of PSD matrices used in existing DR approaches, the assignments provide a hard clustering of the input samples. We discuss key properties advocating for the use of this clustering in Section 2 before introducing our model in Section 3 and applying it to real data in Section 4.

2 Generalization of Dimension Reduction via Graph Matching

Unified view of Dimensionality Reduction. Let $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_N)^\top \in \mathbb{R}^{N\times p}$ be an input dataset of interest. DR methods focus on constructing a low-dimensional representation or *embedding* $\boldsymbol{Z} \in \mathbb{R}^{N\times d}$, where d is smaller than p. The latter should preserve a prescribed geometry for the dataset usually encoded via a pairwise similarity matrix $\boldsymbol{C}_X$. To this end, most popular DR methods (*e.g.* kernel PCA [31], MDS [36], Laplacian eigenmaps [2], SNE-like methods [16]) optimize $\boldsymbol{Z}$ such that its similarity matrix $\boldsymbol{C}_Z$ matches $\boldsymbol{C}_X$ in accordance with the following objective:

$$\mathcal{J}_L(\boldsymbol{C}_X, \boldsymbol{C}_Z) \coloneqq \sum_{(i,j)\in[\![N]\!]^2} L([\boldsymbol{C}_X]_{ij}, [\boldsymbol{C}_Z]_{ij}) \tag{1}$$

where $L : \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+$ is typically the quadratic loss $L_2(x, y) \coloneqq (x - y)^2$ or the generalized Kullback-Leibler divergence $L_{\mathrm{KL}}(x, y) \coloneqq x\log(x/y) - x + y$. As detailed in Appendix A, the definitions of $\boldsymbol{C}_X$ and $\boldsymbol{C}_Z$ as well as L are what differentiate each method. Note that these objectives can be derived from a common Markov random field model with various graph priors [38]. The unified objective Equation (1) can also be seen as a trivial instance of graph matching where both graph structures $\boldsymbol{C}_Z$ and $\boldsymbol{C}_X$ are designed so that their nodes are aligned. To promote clustering from this objective, one can enforce $\boldsymbol{C}_Z$ to have fewer nodes than $\boldsymbol{C}_X$ ($n < N$) and seek for meaningful structural correspondences between the nodes of both graphs.

Gromov-Wasserstein framework. Interestingly, the Optimal Transport (OT, [42, 26]) literature provides a way to do so with the Gromov-Wasserstein discrepancy (GW, [23, 33, 27, 8]). In this context, nodes are endowed with probability weights $\boldsymbol{h}_X \in \Sigma_N$ and $\boldsymbol{h}_Z \in \Sigma_n$ encoding their relative importance. GW then computes a soft-assignment matrix between the nodes of the two graphs $(\boldsymbol{C}_X, \boldsymbol{h}_X)$ and $(\boldsymbol{C}_Z, \boldsymbol{h}_Z)$, as well as a notion of dissimilarity between them reading as:

$$\mathrm{GW}_L(\boldsymbol{C}_X, \boldsymbol{h}_X, \boldsymbol{C}_Z, \boldsymbol{h}_Z) \coloneqq \min_{\boldsymbol{T}\in\mathcal{U}(\boldsymbol{h}_X,\boldsymbol{h}_Z)} \sum_{(i,j)\in[\![N]\!]^2} \sum_{(k,l)\in[\![n]\!]^2} L([\boldsymbol{C}_X]_{ij}, [\boldsymbol{C}_Z]_{kl}) T_{ik} T_{jl} \tag{2}$$

where $\mathcal{U}(\boldsymbol{h}_X, \boldsymbol{h}_Z) = \left\{\boldsymbol{T} \in \mathbb{R}_+^{N\times n} | \boldsymbol{T}\mathbf{1}_n = \boldsymbol{h}_X, \boldsymbol{T}^\top\mathbf{1}_N = \boldsymbol{h}_Z\right\}$. An optimal coupling $\boldsymbol{T}^*$ acts as a soft matching of the nodes, which tends to associate pairs of nodes that have similar pairwise relations in $\boldsymbol{C}_X$ and $\boldsymbol{C}_Z$ respectively. These properties are clear benefits for many ML tasks such as alignments of diverse structured objects [32, 1, 48, 11, 3], (co-)clustering [27, 35], graph representation learning [49, 45, 19, 44, 50] and partitioning [47, 9]. The latter is in line with our objectives as it focuses on the design of a target graph $(\overline{\boldsymbol{C}}, \overline{\boldsymbol{h}})$, so that the OT resulting from $\mathrm{GW}(\boldsymbol{C}_X, \boldsymbol{h}_X, \overline{\boldsymbol{C}}, \overline{\boldsymbol{h}})$ provides a most significant clustering of the nodes in $(\boldsymbol{C}_X, \boldsymbol{h}_X)$. A first axiom consisted in fixing $\overline{\boldsymbol{C}}$ and optimizing its nodes' relative importance $\overline{\boldsymbol{h}}$ modeling cluster proportions [43]. This problem is efficiently tackled using the semi-relaxed GW divergence (srGW) which interest boils down to minimizing the GW loss in Equation (2) over $\mathcal{U}_n(\boldsymbol{h}_X) = \left\{\boldsymbol{T} \in \mathbb{R}_+^{N\times n} | \boldsymbol{T}\mathbf{1}_n = \boldsymbol{h}_X\right\}$. We argue that a better approach consists of also learning the target structure so that its entries would describe connectivity between clusters allowing a sharper graph partitioning. Which leads to the following optimization problem:

$$\min_{\overline{\boldsymbol{C}}\in\mathbb{R}^{n\times n}} \mathrm{srGW}_L(\boldsymbol{C}_X, \boldsymbol{h}_X, \overline{\boldsymbol{C}}) \Leftrightarrow \min_{\overline{\boldsymbol{C}}\in\mathbb{R}^{n\times n}, \overline{\boldsymbol{h}}\in\Sigma_n} \mathrm{GW}_L(\boldsymbol{C}_X, \boldsymbol{h}_X, \overline{\boldsymbol{C}}, \overline{\boldsymbol{h}})\,. \tag{srGWB}$$

This amounts to searching for the closest graph $(\overline{\boldsymbol{C}}, \overline{\boldsymbol{h}})$ of size n to the input graph $(\boldsymbol{C}_X, \boldsymbol{h}_X)$ in the GW sense. As such, it is a specific instance of srGW barycenter over a single input graph [43]. We next study whether srGWB admits OT which are actual membership matrices (with a single non null value per row) achieving hard clusterings of the nodes of $\boldsymbol{C}_X$.

Theorem 1. *Let $\boldsymbol{C}_X \in \mathbb{R}^{N\times N}$ and $\boldsymbol{h}_X \in \Sigma_N^*$ a vector in the probability simplex. If $g(\boldsymbol{U}) = \mathrm{vec}(\boldsymbol{U})^\top (\boldsymbol{C}_X \otimes_K \boldsymbol{C}_X)\,\mathrm{vec}(\boldsymbol{U})$ is convex on $\mathcal{U}(\boldsymbol{h}_X, \boldsymbol{h}_X)$, then srGWB with $L = L_2$ admits scaled membership matrices as optimum.*

The sufficient condition in Theorem 1 is satisfied for existing DR methods (Appendix A), *e.g* when $\boldsymbol{C}_X$ is PSD (or NSD). In this setting, this result completes the analysis of [7] establishing that srGWB constrained to membership matrices as OT is a SOTA graph coarsening method for spectrum preservation, equivalent to a weighted kernel K-means [12, 13]. Following [41, equation 6],we can also see that g is convex whenever the GW problem from a graph to itself is concave. Hence [29, Proposition 2] also extends our analysis to squared Euclidean distance matrices. A corollary of Theorem 1 establishes an analog result when $\overline{\boldsymbol{h}}$ is not optimized (Appendix B).

3 Joint Clustering and Dimensionality Reduction

Dimensionality reduction with Gromov-Wasserstein. In light of the results presented above on the clustering abilities of srGW, we introduce a versatile algorithm for joint clustering and dimensionality reduction. Our method amounts to replacing the usual DR objective Equation (1) by a srGW loss Equation (2) thus allowing to reduce the sample size. Namely, we learn embeddings $\boldsymbol{Z}$ that parametrize a structure $\boldsymbol{C}_Z$ induced by the underlying DR method as follows:

$$\min_{\boldsymbol{Z}\in\mathbb{R}^{n\times d}} \mathrm{srGW}_L(\boldsymbol{C}_X, \boldsymbol{h}_X, \boldsymbol{C}_Z)\,. \tag{GW-DR}$$

The embeddings $\boldsymbol{Z}$ then act as low-dimensional prototypical representations of input samples, whose learned relative importance $\boldsymbol{h}_Z$ accommodates clusters or substructures of varying proportions in $\boldsymbol{X}$. When $\boldsymbol{C}_Z = \boldsymbol{Z}\boldsymbol{Z}^\top$ mimicking e.g PCA (Appendix A), GW-DR boils down to a srGW barycenter problem constrained to have at most rank d which coincides with srGWB if $d \geq n$. These relations and Theorem 1 allow us to expect OT solutions close to providing a hard-clustering of $\boldsymbol{X}$. Finally, we emphasize that the GW framework does not take into account input samples and embeddings explicitly, but only implicitly through their pairwise similarity matrices $\boldsymbol{C}_X$ and $\boldsymbol{C}_Z$. To readily incorporate the feature information of $\boldsymbol{X}$ in GW-DR, one can adopt the Fused GW framework [34] that interpolates linearly, via a hyperparameter $\alpha \in [0, 1]$, between our objective and a linear OT cost that matches samples $\boldsymbol{X} \in \mathbb{R}^{N\times p}$ and a learned feature matrix $\overline{\boldsymbol{F}} \in \mathbb{R}^{n\times p}$. The latter essentially reduces to a concave problem, wherein the goal is to achieve K-means clustering on $\boldsymbol{X}$ [4], hence acting as a concave regularization of GW-DR (see details in Appendix C.1).

Computation. GW-DR is a non-convex problem that we propose to tackle using a Block Coordinate Descent algorithm (BCD, [37]) guaranteed to converge to local optimum [15, 20]. The BCD alternates between *i)* solving for a srGW problem given $\boldsymbol{Z}$ using the Conditional Gradient solver in [43] extended to support L_{KL}; *ii)* optimizing $\boldsymbol{Z}$ for a fixed OT using gradient descent with adaptive learning rates [17]. Each update is achieved in $\mathcal{O}(nN^2 + n^2N)$ operations.

Related work. The closest to our work is the COOT-clustering approach proposed in [29] that estimates simultaneously a clustering of samples and variables using the CO-Optimal Transport problem. The key difference is that we leverage the affinity matrices and kernels of existing DR methods instead of aligning the features. Other approaches such as [18] involve modelling latent variables with mixture distributions. Note that none of the previously proposed methods can easily adapt to the mechanisms of existing DR methods like Equation (GW-DR).

4 Experiments

In this section, we showcase the relevance of our approach on popular image datasets: COIL-20 [25], MNIST and fashion-MNIST [46]. Results are averages and standard deviations, computed over 5 runs with different random seeds. Details about evaluation metrics and datasets are provided in Appendix C. Throughout this section, we set $\boldsymbol{h}_X$ as uniform. In what follows, for any existing DR method, we refer to its gromovized version by appending the prefix "GW" to the method name *e.g.* GW-PCA.

Table 1: ARI (%) clustering scores.

	srGWI	srGWB
MNIST	29.7(1.9)	**32.6**(**1.8**)
F-MNIST	26.1(0.0)	**39.5**(**0.3**)
COIL	18.1(0.2)	**51.0**(**1.7**)

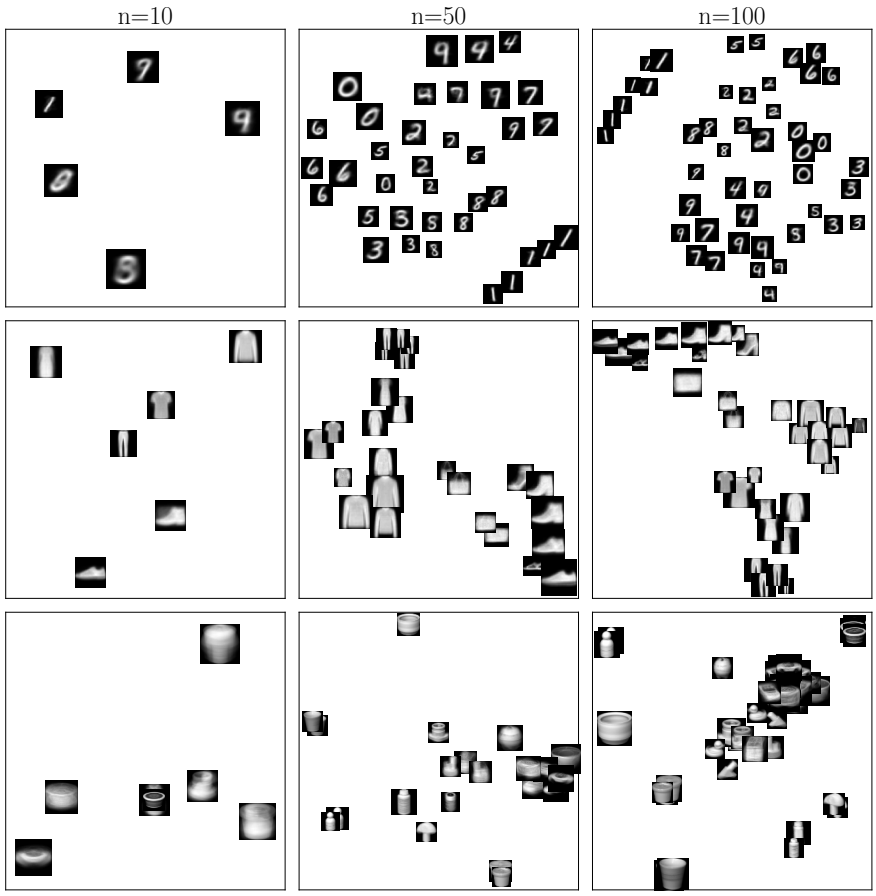

Figure 1: 2D Embeddings of GW-tSNEkhorn applied to MNIST (top), Fashion-MNIST (middle) and COIL (bottom) with various n. The perplexity is set to $\xi = 50$ for all experiments. Images for prototypes are computed as Wasserstein barycenters of the associated input images. Their areas are proportional to $\boldsymbol{h}_Z$.

Clustering. We first evaluate the clustering abilities of srGW barycenters (srGWB) and their vanilla counterpart with fixed structure $\boldsymbol{I}_n$ used in the graph partitioning literature (srGWI, [43]). For both, $\boldsymbol{C}_X$ is taken as the MDS kernel (see Appendix A). Clustering performances measured by means of ARI are reported in Table 1 and show the superiority of srGWB.

Joint Clustering and Dimensionality Reduction. In Figure 1, we display the prototypes produced by GW-tSNEkhorn (robust version of tSNE presented in [39]) for various n. We used fused srGW [41] with $\alpha = 0.5$ as it naturally produces prototypes in input space (as Wasserstein barycenters of images) that can be visualized. They show the relatively effective purity of the prototypes confirmed by the homogeneity scores displayed in Table 2 for various n. Recall that n only provides an upper bound of the number of prototypes as the semi-relaxed OT problem permits the flexibility to discard unnecessary prototypes. The latter scores compute to which extent prototypes contain samples of the same label. It's reasonable to note that as the value of n increases, the consistency or similarity among the prototypes also increases.

Table 2: Homogeneity ($\times 100$) scores for GW-tSNEkhorn.

	$n = 10$	$n = 50$	$n = 100$	$n = 200$
MNIST	49.2(1.5)	76.8(1.1)	80.8(0.6)	83.8(0.8)
F-MNIST	56.0(2.4)	68.9(0.7)	69.8(1.4)	71.9(1.6)
COIL	55.8(1.2)	77.9(3.2)	82.2(3.6)	85.3(2.9)

Should clustering depend on embeddings? Choosing the fused GW hyperparameter as $\alpha \to 0$ would result in the clustering ignoring the current positions of embeddings and only leveraging information about the input $\boldsymbol{X}$ (pure clustering). To determine whether this can be beneficial, we performed a grid search over different values of α (details in Appendix C). We selected the value

α^* that maximizes the sum of homogeneity and silhouette scores [30]. The latter is computed based on a ground truth taken as the most represented input label in the associated prototype. Thus it gives a quantitative metric to properly evaluate the prototypes' relative positions. Best scores and their respective α^* are reported in Table 3 for GW-tSNEkhorn. These illustrate the significance of embedding-dependent clustering to ensure that the embeddings display a meaningful structure, as all α^* are greater than 0.

5 Concluding Remarks

We believe that the versatility of our approach will enable applications beyond data visualization. For instance, the formalism associated with (sr)GW barycenters naturally allows us to consider multiple affinity matrices as inputs. In this context, popular open challenges relate to the multi-scale and multi-view dimensionality reduction problems. We envision to thoroughly investigate the latter both empirically and theoretically, building on Theorem 1 which may also conduct to new discoveries for the GW-based (multi) graph coarsening or dictionary learning.

Table 3: Best fused GW parameter α for $n = 50$.

	α^*	Homogeneity	Silhouette
MNIST	0.9997	74.7(0.2)	16.4(5.6)
F-MNIST	1	61.5(1.6)	16.3(3.6)
COIL	0.999999	87.51(0.1)	42.1(5.6)

References

[1] David Alvarez-Melis and Tommi S Jaakkola. Gromov-wasserstein alignment of word embedding spaces. *arXiv preprint arXiv:1809.00013*, 2018.

[2] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[3] Clément Bonet, Titouan Vayer, Nicolas Courty, François Septier, and Lucas Drumetz. Subspace detours meet gromov–wasserstein. *Algorithms*, 14(12):366, 2021.

[4] Guillermo Canas and Lorenzo Rosasco. Learning probability measures with respect to optimal transport metrics. *Advances in Neural Information Processing Systems*, 25, 2012.

[5] Laura Cantini, Pooya Zakeri, Celine Hernandez, Aurelien Naldi, Denis Thieffry, Elisabeth Remy, and Anaïs Baudot. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature communications*, 12(1):124, 2021.

[6] Lei Cao, Darian McLaren, and Sarah Plosker. Centrosymmetric stochastic matrices. *Linear and Multilinear Algebra*, 70(3):449–464, 2022.

[7] Yifan Chen, Rentian Yao, Yun Yang, and Jie Chen. A gromov–wasserstein geometric view of spectrum-preserving graph coarsening. *arXiv preprint arXiv:2306.08854*, 2023.

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

[9] Samir Chowdhury and Tom Needham. Generalized spectral clustering via gromov-wasserstein learning. In *International Conference on Artificial Intelligence and Statistics*, pages 712–720. PMLR, 2021.

[10] Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

[11] Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. Gromov-wasserstein optimal transport to align single-cell multi-omics data. *BioRxiv*, pages 2020–04, 2020.

[12] Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–556, 2004.

[13] Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Weighted graph cuts without eigenvectors a multilevel approach. *IEEE transactions on pattern analysis and machine intelligence*, 29(11):1944–1957, 2007.

[14] David L Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1(2000):32, 2000.

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

[16] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] Wei Liu, Xu Liao, Yi Yang, Huazhen Lin, Joe Yeong, Xiang Zhou, Xingjie Shi, and Jin Liu. Joint dimension reduction and clustering analysis of single-cell rna-seq and spatial transcriptomics data. *Nucleic acids research*, 50(12):e72–e72, 2022.

[19] Weijie Liu, Jiahao Xie, Chao Zhang, Makoto Yamada, Nenggan Zheng, and Hui Qian. Robust graph dictionary learning. In *The Eleventh International Conference on Learning Representations*, 2022.

[20] Hanbaek Lyu and Yuchen Li. Block majorization-minimization with diminishing radius for constrained nonconvex optimization. 08 2023.

[21] KV Mardia, JT Kent, and JM Bibby. Multivariate analysis, 1979. *Probability and mathematical statistics. Academic Press Inc*, 1979.

[22] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

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

[24] Ariana Mendible, Steven L Brunton, Aleksandr Y Aravkin, Wes Lowrie, and J Nathan Kutz. Dimensionality reduction and reduced-order modeling for traveling wave physics. *Theoretical and Computational Fluid Dynamics*, 34:385–400, 2020.

[25] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-20). 1996.

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

[27] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pages 2664–2672. PMLR, 2016.

[28] Nathalie Pochet, Frank De Smet, Johan AK Suykens, and Bart LR De Moor. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics*, 20(17):3185–3195, 2004.

[29] Ievgen Redko, Titouan Vayer, Rémi Flamary, and Nicolas Courty. Co-optimal transport. *Advances in Neural Information Processing Systems*, 33(17559-17570):2, 2020.

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

[31] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer, 1997.

[32] Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016.

[33] Karl-Theodor Sturm. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. *arXiv preprint arXiv:1208.0434*, 2012.

[34] Vayer Titouan, Nicolas Courty, Romain Tavenard, Chapel Laetitia, and Rémi Flamary. Optimal transport for structured data with application on graphs. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6275–6284. PMLR, 09–15 Jun 2019.

[35] Vayer Titouan, Ievgen Redko, Rémi Flamary, and Nicolas Courty. Co-optimal transport. *Advances in neural information processing systems*, 33:17559–17570, 2020.

[36] Warren S Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

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

[38] Hugues Van Assel, Thibault Espinasse, Julien Chiquet, and Franck Picard. A probabilistic graph coupling view of dimension reduction. *Advances in Neural Information Processing Systems*, 35:10696–10708, 2022.

[39] Hugues Van Assel, Titouan Vayer, Rémi Flamary, and Nicolas Courty. Snekhorn: Dimension reduction with symmetric entropic affinities. *arXiv preprint arXiv:2305.13797*, 2023.

[40] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[41] Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Optimal transport for structured data with application on graphs. *arXiv preprint arXiv:1805.09114*, 2018.

[42] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

[43] Cédric Vincent-Cuaz, Rémi Flamary, Marco Corneli, Titouan Vayer, and Nicolas Courty. Semi-relaxed gromov-wasserstein divergence with applications on graphs. *arXiv preprint arXiv:2110.02753*, 2021.

[44] Cédric Vincent-Cuaz, Rémi Flamary, Marco Corneli, Titouan Vayer, and Nicolas Courty. Template based graph neural network with optimal transport distances. *Advances in Neural Information Processing Systems*, 35:11800–11814, 2022.

[45] Cédric Vincent-Cuaz, Titouan Vayer, Rémi Flamary, Marco Corneli, and Nicolas Courty. Online graph dictionary learning. In *International conference on machine learning*, pages 10564–10574. PMLR, 2021.

[46] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[47] Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable gromov-wasserstein learning for graph partitioning and matching. *Advances in neural information processing systems*, 32, 2019.

[48] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. Gromov-wasserstein learning for graph matching and node embedding. In *International conference on machine learning*, pages 6932–6941. PMLR, 2019.

[49] Hongtengl Xu. Gromov-wasserstein factorization models for graph clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6478–6485, 2020.

[50] Zhichen Zeng, Ruike Zhu, Yinglong Xia, Hanqing Zeng, and Hanghang Tong. Generative graph dictionary learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 40749–40769. PMLR, 23–29 Jul 2023.

A Framing Dimensionality Reduction as Graph Matching

In this section, we provide a unified view of the most popular DR methods with the following objective, where $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is the loss function,

$$\mathcal{J}_L(\boldsymbol{C}_X, \boldsymbol{C}_Z) \coloneqq \sum_{(i,j)\in[\![N]\!]^2} L([\boldsymbol{C}_X]_{ij}, [\boldsymbol{C}_Z]_{ij}) \,. \tag{3}$$

Kernel PCA, MDS and Isomap. Let us consider $\boldsymbol{K}_X = (k(\boldsymbol{x}_i, \boldsymbol{x}_j))_{ij} \in \mathcal{S}_+^N$ a kernel matrix over the input data $\boldsymbol{X}$. Denoting $R_d \coloneqq \{\boldsymbol{C} \in \mathcal{S}_+^N \text{ s.t. } \mathrm{rk}(\boldsymbol{C}) \leq d\}$ the set of rank at most d PSD matrices, kernel PCA [31] computes $\boldsymbol{S}_Z = \mathrm{Proj}_{R_d}^F(\boldsymbol{K}_X)$. Since $\boldsymbol{S}_Z \in R_d$, we have the existence of $\boldsymbol{Z} \in \mathbb{R}^{N\times d}$ such that $\boldsymbol{S}_Z = \boldsymbol{Z}\boldsymbol{Z}^\top$ (sample covariance of $\boldsymbol{Z}$). In view of this property, the kernel PCA problem reads

$$\min_{\boldsymbol{Z}\in\mathbb{R}^{n\times d}} \mathcal{J}_{L_2}(\boldsymbol{K}_X, \boldsymbol{S}_Z) \,. \tag{PCA}$$

Note that traditional PCA simply amounts to choosing $\boldsymbol{K}_X = \boldsymbol{X}\boldsymbol{X}^\top$ in the above problem. Multidimensional scaling (MDS) [36] can be easily derived from a slight variation of PCA. Define $\boldsymbol{D}_X = -\boldsymbol{H}_N\boldsymbol{E}_X\boldsymbol{H}_N$ with $[\boldsymbol{E}_X]_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$ and where $\boldsymbol{H}_N = \boldsymbol{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N^\top$ is the centering matrix. Since $\boldsymbol{E}_X$ is a squared Euclidean distance matrix, it results that $\boldsymbol{D}_X \in \mathcal{S}_+^N$ [21]. Classical MDS then amounts to minimizing the following strain.

$$\min_{\boldsymbol{Z}\in\mathbb{R}^{n\times d}} \mathcal{J}_{L_2}(\boldsymbol{D}_X, \boldsymbol{S}_Z) \,. \tag{MDS}$$

Laplacian Eigenmaps. Let $\boldsymbol{W}_X \in \mathcal{S}^N$ be a similarity graph (*e.g.* neighborhood graph) built from $\boldsymbol{X}$. We define its graph Laplacian as $\boldsymbol{L}_X = \mathrm{diag}(\boldsymbol{W}_X\mathbf{1}) - \boldsymbol{W}_X$ such that $\boldsymbol{L}_X \in \mathcal{S}_+^N$ [10]. Laplacian eigenmaps [2] boils down to the following objective

$$\max_{\boldsymbol{Z}\in\mathrm{St}(n,d)} \mathcal{J}_{L_2}(\boldsymbol{L}_X, \boldsymbol{S}_Z) \tag{LE}$$

where $\mathrm{St}(n,d) = \{\boldsymbol{U} \in \mathbb{R}^{n\times d}, \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_d\}$ is the orthogonal Stiefel manifold. This constraint prevents the embeddings from collapsing to $\mathbf{0}$.

Neighbor Embedding. Another popular class of methods is the neighbor embedding framework. The central idea is to minimize the Kullback-Leibler divergence between two kernels $\boldsymbol{K}_X$ and $\boldsymbol{K}_Z$.

$$\min_{\boldsymbol{Z}\in\mathbb{R}^{n\times d}} \mathcal{J}_{L_{\mathrm{KL}}}(\boldsymbol{K}_X, \boldsymbol{K}_Z) \tag{NE}$$

Although some methods leave the kernels unnormalized (*e.g.* UMAP by [22]), the latter are usually taken as either row-stochastic (*e.g.* SNE by [16] and t-SNE by [40]) or doubly-stochastic normalized (SNEkhorn by [39]). We briefly detail the latter as we rely on it in our experiments in Section 4. It consists in controlling the entropy in each point by solving the following OT problem

$$\min_{\boldsymbol{P}\in\mathcal{H}_\xi\cap\mathcal{S}} \langle \boldsymbol{P}, \boldsymbol{C}\rangle \,. \tag{4}$$

with $\mathcal{H}_\xi \coloneqq \{\boldsymbol{P} \in \mathbb{R}_+^{n\times n} \text{ s.t. } \boldsymbol{P}\mathbf{1} = \mathbf{1} \text{ and } \forall i,\ \mathrm{H}(\boldsymbol{P}_{i:}) \geq \log\xi + 1\}$ where the entropy of $\boldsymbol{p} \in \mathbb{R}_+^n$ is[1] $\mathrm{H}(\boldsymbol{p}) = -\sum_i p_i(\log(p_i) - 1) = -\langle \boldsymbol{p}, \log\boldsymbol{p} - \mathbf{1}\rangle$. Note that at the optimum the entropy constraint is saturated thus allowing to accommodate for potentially varying noise levels while producing a doubly stochastic symmetric affinity matrix.

B (Semi-relaxed) Gromov-Wasserstein barycenter as a concave OT problem

We consider here any graph $\mathcal{G} = (\boldsymbol{C}, \boldsymbol{h})$ modeled as a connectivity matrix $\boldsymbol{C} \in \mathbb{R}^{N\times N}$ and $\overline{\boldsymbol{C}} \in \mathbb{R}^{n\times n}$ and a probability vector $\boldsymbol{h} \in \Sigma_N^*$ and $\overline{\boldsymbol{h}} \in \Sigma_n^*$. We focus next on the semi-relaxed Gromov-Wasserstein barycenter problem with an euclidean inner cost ($L = L_2$) reading as follows

$$\min_{\overline{\boldsymbol{C}}\in\mathbb{R}^{n\times n}} \mathrm{srGW}(\boldsymbol{C}, \boldsymbol{h}, \overline{\boldsymbol{C}}) \quad \Leftrightarrow \quad \min_{\overline{\boldsymbol{C}}\in\mathbb{R}^{n\times n}} \min_{\boldsymbol{T}\in\mathcal{U}_n(\boldsymbol{h})} \mathcal{E}_2(\boldsymbol{C}, \overline{\boldsymbol{C}}, \boldsymbol{T}) \tag{srGW-bary1}$$

[1] With the convention $0\log 0 = 0$.

where $\mathcal{E}_2$ coincides with the objective function in equation 2 applied in $\boldsymbol{C}$ and $\overline{\boldsymbol{C}}$. In general, the latter is considered as a non-convex problem. Notice that the subproblem w.r.t $\overline{\boldsymbol{C}}$ is convex. While the subproblem w.r.t $\boldsymbol{T}$ is in general non-convex and is equivalent to a quadractic program with Hessian matrix $\mathcal{H} = \overline{\boldsymbol{C}}^2 \otimes_K \mathbf{1}_N \mathbf{1}_N^\top - 2\overline{\boldsymbol{C}} \otimes_K \boldsymbol{C}$, where $\otimes_K$ is the kronecker product and the power operation is taken element-wise.

In the following, we proof a sufficient condition so that membership matrices are optimal for the srGW-bary1 problem stated as such:

Theorem 1. *Let $\boldsymbol{C} \in \mathbb{R}^{N\times N}$ any bounded matrix and $\boldsymbol{h} \in \Sigma_N^*$. Every solutions to the following problem*

$$\min_{\boldsymbol{T}\in\mathcal{U}_n(\boldsymbol{h})} \mathcal{E}_2(\boldsymbol{C}, \widetilde{\boldsymbol{C}}(\boldsymbol{T}), \boldsymbol{T}) \;\; \textit{with} \;\; \forall (i,j) \in [\![n]\!]^2 \;,\; \overline{\boldsymbol{C}}(\boldsymbol{T})_{ij} = \begin{cases} \left(\boldsymbol{T}^\top \boldsymbol{C}\boldsymbol{T} \oslash \overline{\boldsymbol{h}}\,\overline{\boldsymbol{h}}^\top\right)_{ij} & \textit{if } \overline{h}_i\overline{h}_j > 0 \\ 0 \;\; \textit{otherwise.} \end{cases} \tag{srGW-bary2}$$

and $\overline{\boldsymbol{h}} = \boldsymbol{T}^\top \mathbf{1}_N$ are solutions to the srGW-bary1 problem. Moreover If the function $g_{\boldsymbol{C}}$ defined for any $\boldsymbol{U} \in \mathcal{U}(\boldsymbol{h}, \boldsymbol{h})$ as

$$g_{\boldsymbol{C}}(\boldsymbol{U}) = \operatorname{vec}(\boldsymbol{U})^\top \left(\boldsymbol{C} \otimes_K \boldsymbol{C}\right) \operatorname{vec}(\boldsymbol{U}) \tag{5}$$

is <u>convex</u> on $\mathcal{U}(\boldsymbol{h}, \boldsymbol{h})$, then the srGW-bary2 problem is <u>concave</u> on $\mathcal{U}_n(\boldsymbol{h})$, hence problem srGW-bary1 admits extremities of $\mathcal{U}_n(\boldsymbol{h})$ as OT solutions.

To prove Theorem 1, let us begin with proving the following Lemma:

Lemma 1. *For any bounded matrix $\boldsymbol{C} \in \mathbb{R}^{N\times N}$ and probability vector $\boldsymbol{h} \in \Sigma_N^*$, every solutions to problem srGW-bary2 are solutions to problem srGW-bary1.*

Proof of Lemma 1. Let us first characterize solutions to the srGW-bary1 problem. Let $\boldsymbol{T} \in \mathcal{U}_n(\boldsymbol{h})$. We will find a minimizer of the convex function $\overline{\boldsymbol{C}} \in \mathbb{R}^{n\times n} \mapsto \mathcal{E}_2(\boldsymbol{C}, \overline{\boldsymbol{C}}, \boldsymbol{T})$ that we will denote by $\overline{\boldsymbol{C}}$. Using the first order conditions and the convexity of this function, $\overline{\boldsymbol{C}}$ is a solution *if and only if* it satisfies

$$\nabla_{\overline{\boldsymbol{C}}} \mathcal{E}_2(\boldsymbol{C}, \overline{\boldsymbol{C}}, \boldsymbol{T}) = 2\overline{\boldsymbol{C}} \odot \overline{\boldsymbol{h}}\,\overline{\boldsymbol{h}}^\top - 2\boldsymbol{T}^\top \boldsymbol{C}\boldsymbol{T} = \mathbf{0}\,, \tag{6}$$

where $\overline{\boldsymbol{h}}$ depends on $\boldsymbol{T}$ and is defined as $\overline{\boldsymbol{h}} = \boldsymbol{T}^\top \mathbf{1}_N = \left(\|\boldsymbol{T}_{:,1}\|_1, \cdots, \|\boldsymbol{T}_{:,n}\|_1\right)^\top$ and $\boldsymbol{T}_{:,j} \in \mathbb{R}^N$ denotes the jth column of $\boldsymbol{T}$. We define

$$\forall (i,j) \in [\![n]\!]^2,\; \overline{\boldsymbol{C}}(\boldsymbol{T})_{ij} = \begin{cases} \left(\boldsymbol{T}^\top \boldsymbol{C}\boldsymbol{T} \oslash \overline{\boldsymbol{h}}\,\overline{\boldsymbol{h}}^\top\right)_{ij} & \text{if } \overline{h}_i\overline{h}_j = \|\boldsymbol{T}_{:,i}\|_1 \|\boldsymbol{T}_{:,j}\|_1 > 0 \\ 0 \text{ otherwise.} \end{cases} \tag{7}$$

On one hand for $(i,j) \in [\![n]\!]^2$ such that $\overline{h}_i\overline{h}_j > 0$, $\overline{\boldsymbol{C}}(\boldsymbol{T})_{ij} = \left(\boldsymbol{T}^\top \boldsymbol{C}\boldsymbol{T} \oslash \overline{\boldsymbol{h}}\,\overline{\boldsymbol{h}}^\top\right)_{ij}$ which clearly satisfies the first order conditions Equation (6). On the other hand, for $(i,j) \in [\![n]\!]^2$ such that $\overline{h}_i\overline{h}_j = 0$, we have $\boldsymbol{T}_{:,i} = \mathbf{0}$ or $\boldsymbol{T}_{:,j} = \mathbf{0}$, as $\forall i, j,\; T_{ij} \geq 0$. Hence

$$(2\overline{\boldsymbol{C}}(\boldsymbol{T}) \odot \overline{\boldsymbol{h}}\,\overline{\boldsymbol{h}}^\top - 2\boldsymbol{T}^\top \boldsymbol{C}\boldsymbol{T})_{ij} = (-2\boldsymbol{T}^\top \boldsymbol{C}\boldsymbol{T})_{ij} = -2\boldsymbol{T}_{:,i}^\top \boldsymbol{C}\boldsymbol{T}_{:,j} = 0 \tag{8}$$

Overall, $\overline{\boldsymbol{C}}(\boldsymbol{T})$ satisfies the first order conditions and thus is minimizing $\overline{\boldsymbol{C}} \in \mathbb{R}^{n\times n} \mapsto \mathcal{E}_2(\boldsymbol{C}, \overline{\boldsymbol{C}}, \boldsymbol{T})$. Consequently solutions to Equation (srGW-bary1) can be found by minimizing

$$\mathcal{F} : \boldsymbol{T} \in \mathcal{U}_n(\boldsymbol{h}) \mapsto \mathcal{E}_2(\boldsymbol{C}, \overline{\boldsymbol{C}}(\boldsymbol{T}), \boldsymbol{T}) = \sum_{ijkl} |C_{ij} - \overline{\boldsymbol{C}}(\boldsymbol{T})_{kl}|^2 T_{ik} T_{jl} \tag{9}$$

In order to prove the existence of a minimizer of $\mathcal{F}$ we will show that it is continuous on $\mathcal{U}_n(\boldsymbol{h})$ and conclude by compactness of $\mathcal{U}_n(\boldsymbol{h})$. For any $\boldsymbol{T} \in \mathcal{U}_n(\boldsymbol{h})$, we have

$$\mathcal{F}(\boldsymbol{T}) = \sum_{ij} C_{ij}^2 h_i h_j + \sum_{kl} \overline{\boldsymbol{C}}(\boldsymbol{T})_{kl}^2 \overline{h}_k \overline{h}_l - 2\sum_{ijkl} C_{ij} \overline{\boldsymbol{C}}(\boldsymbol{T})_{kl} T_{ik} T_{jl} \tag{10}$$

Now by definition of $\overline{\boldsymbol{C}}(\boldsymbol{T})$ it satisfies the first order conditions Equation (6) and in particular

$$\forall (k,l) \in [\![n]\!]^2,\; \overline{\boldsymbol{C}}(\boldsymbol{T})_{ij} \overline{h}_k \overline{h}_l = \boldsymbol{T}_{:,k}^\top \boldsymbol{C}\boldsymbol{T}_{:,l}\,. \tag{11}$$

Thus $\sum_{kl} \overline{\boldsymbol{C}}(\boldsymbol{T})_{kl}^2 \overline{h}_k \overline{h}_l = \sum_{kl} \overline{\boldsymbol{C}}(\boldsymbol{T})_{kl} \boldsymbol{T}_{:,k}^\top \boldsymbol{C} \boldsymbol{T}_{:,l} = \sum_{ijkl} C_{ij} \overline{\boldsymbol{C}}(\boldsymbol{T})_{kl} T_{ik} T_{jl}$. Consequently the two last terms of $\mathcal{F}(\boldsymbol{T})$ simplify and we can reformulate

$$\begin{aligned} \mathcal{F}(\boldsymbol{T}) &= \sum_{ij} C_{ij}^2 h_i h_j - \sum_{kl} \overline{\boldsymbol{C}}(\boldsymbol{T})_{kl} \boldsymbol{T}_{:,k}^\top \boldsymbol{C} \boldsymbol{T}_{:,l} \\ &= \sum_{ij} C_{ij}^2 h_i h_j - \sum_{\substack{k:\overline{h}_k \neq 0 \\ l:\overline{h}_l \neq 0}} \frac{(\boldsymbol{T}_{:,k}^\top \boldsymbol{C} \boldsymbol{T}_{:,l})^2}{\overline{h}_k \overline{h}_l} \end{aligned} \tag{12}$$

which is continuous on $\mathcal{U}_n(\boldsymbol{h})$.

Therefore we ensured that srGW-bary1 admits solutions of the form $(\boldsymbol{T}, \overline{\boldsymbol{C}}(\boldsymbol{T}))$ where $\overline{\boldsymbol{C}}(\boldsymbol{T})$ satisfies equation 7. Moreover, these solutions can be found by minimizing w.r.t $\boldsymbol{T} \in \mathcal{U}_n(\boldsymbol{h})$ the function $\mathcal{F}$ defined in equation 9, which coincides with the problem srGW-bary2.

□

Concavity analysis. The proof of Theorem 1 consists in studying the concavity on $\mathcal{U}_n(\boldsymbol{h})$ of the objective function $\mathcal{F}$ involved in problem srGW-bary2. To this end, we will prove that $\mathcal{F}$ is above its tangents. However, we can see from equation 12 that $\mathcal{F}$ is only differentiable on $\mathcal{U}_n(\boldsymbol{h}) \backslash \overset{\circ}{\mathcal{U}}_n(\boldsymbol{h})$, where

$$\overset{\circ}{\mathcal{U}}_n(\boldsymbol{h}) := \left\{ \boldsymbol{T} \in \mathcal{U}_n(\boldsymbol{h}) \mid \boldsymbol{T}^\top \mathbf{1}_N = \overline{\boldsymbol{h}} > \mathbf{0}_n \right\} \tag{13}$$

is a convex subset of $\mathcal{U}_n(\boldsymbol{h})$. As $\mathcal{F}$ reads as a sum of rational functions whose respective denominator $\overline{h}_k \overline{h}_l = 0$ if and only if $\overline{h}_k = 0$ or $\overline{h}_l = 0$. Then we will first study the concavity of $\mathcal{F}$ on $\overset{\circ}{\mathcal{U}}_n(\boldsymbol{h})$. Then we will conclude on the concavity of $\mathcal{F}$ on $\mathcal{U}_n(\boldsymbol{h})$ by an argument of continuity. Notice that the concavity of $\mathcal{F}$ on $\mathcal{U}_n(\boldsymbol{h})$ is equivalent to the convexity of the function

$$f : \boldsymbol{T} \in \mathcal{U}_n(\boldsymbol{h}) \mapsto \sum_{\substack{k:\overline{h}_k \neq 0 \\ l:\overline{h}_l \neq 0}} \frac{(\boldsymbol{T}_{:,k}^\top \boldsymbol{C} \boldsymbol{T}_{:,l})^2}{\overline{h}_k \overline{h}_l} \tag{14}$$

which we will use next for the sake of simplicity. We start by emphasizing in the following lemma a low-rank factorization of f which explicits its link with a GW problem from a graph to itself:

Lemma 2. *$\mathcal{F}$ admits as an equivalent low-rank formulation*

$$\boldsymbol{U} \in \mathcal{V}_n(\boldsymbol{h}) \to g_{\boldsymbol{C}}(\boldsymbol{U}) := \operatorname{vec}(\boldsymbol{U})^\top (\boldsymbol{C} \otimes_K \boldsymbol{C}) \operatorname{vec}(\boldsymbol{U}) \tag{15}$$

where $\mathcal{V}_n(\boldsymbol{h}) := \left\{ \boldsymbol{U} \in \mathbb{R}^{N \times N} | \exists \boldsymbol{T} \in \overset{\circ}{\mathcal{U}}_n(\boldsymbol{h}) \; s.t \; \boldsymbol{T}^\top \mathbf{1}_N = \overline{\boldsymbol{h}}, \; \boldsymbol{U} = \boldsymbol{T} diag(\overline{\boldsymbol{h}}^{-1}) \boldsymbol{T}^\top \right\} \subset \mathcal{U}(\boldsymbol{h}, \boldsymbol{h})$.

Proof of Lemma 2. For any $\boldsymbol{T} \in \overset{\circ}{\mathcal{U}}_n(\boldsymbol{h})$, f can be expressed as

$$\begin{aligned} f(\boldsymbol{T}) &= \| \boldsymbol{D}_{\overline{\boldsymbol{h}}}^{-1/2} \boldsymbol{T}^\top \boldsymbol{C} \boldsymbol{T} \boldsymbol{D}_{\overline{\boldsymbol{h}}}^{-1/2} \|_F^2 \\ &= \operatorname{Tr} \left\{ \boldsymbol{T} \boldsymbol{D}_{\overline{\boldsymbol{h}}}^{-1} \boldsymbol{T}^\top \boldsymbol{C}^\top \boldsymbol{T} \boldsymbol{D}_{\overline{\boldsymbol{h}}}^{-1} \boldsymbol{T}^\top \boldsymbol{C} \right\} \\ (\text{posing } \boldsymbol{U} = \boldsymbol{T} \boldsymbol{D}_{\overline{\boldsymbol{h}}}^{-1} \boldsymbol{T}^\top) \quad &= \operatorname{Tr} \left\{ \boldsymbol{U} \boldsymbol{C}^\top \boldsymbol{U} \boldsymbol{C} \right\} \\ &= \operatorname{vec}(\boldsymbol{U}^\top)^\top (\boldsymbol{C} \otimes_K \boldsymbol{C}) \operatorname{vec}(\boldsymbol{U}) \\ &= \operatorname{vec}(\boldsymbol{U})^\top (\boldsymbol{C} \otimes_K \boldsymbol{C}) \operatorname{vec}(\boldsymbol{U}) := g_{\boldsymbol{C}}(\boldsymbol{U}) \end{aligned} \tag{16}$$

where vec denotes the column stacking operator and $\otimes_K$ the kronecker product. Following e.g [41, equation 6], one can see that $g_{\boldsymbol{C}}$ relates to a low-rank Gromov-Wasserstein problem for a graph $\boldsymbol{C}$ to itself, as $(\boldsymbol{U} = \boldsymbol{T} \boldsymbol{D}_{\overline{\boldsymbol{h}}}^{-1} \boldsymbol{T}^\top)$ is a coupling in $\mathcal{U}(\boldsymbol{h}, \boldsymbol{h})$, resulting from the "self-gluing" of $\boldsymbol{T} \in \overset{\circ}{\mathcal{U}}_n(\boldsymbol{h})$ where $rank(\boldsymbol{T}) \leq n$. □

Then we establish the following result

Lemma 3. *If the function $g_{\boldsymbol{C}}(\boldsymbol{U}) = \mathrm{vec}(\boldsymbol{U})^\top (\boldsymbol{C} \otimes_K \boldsymbol{C}) \mathrm{vec}(\boldsymbol{U})$ is convex on $\mathcal{U}(\boldsymbol{h}, \boldsymbol{h})$, then $\mathcal{F}$ is concave on $\mathring{\mathcal{U}}_n(\boldsymbol{h})$.*

Proof of Lemma 3. To establish the concavity of $\mathcal{F}$ on $\mathring{\mathcal{U}}_n(\boldsymbol{h})$, it suffices to prove that the function f defined in equation 14 is convex on this set. To this end, as f is in $\mathcal{C}^1(\mathring{\mathcal{U}}_n(\boldsymbol{h}), \mathbb{R}_+)$, we will prove that it is above its tangents. For any $(a, b) \in [\![N]\!] \times [\![n]\!]$, its first partial derivates read as

$$
\begin{aligned}
&\frac{\partial}{\partial T_{ab}} f(\boldsymbol{T}) \\
&= \sum_{ij} \left\{ 2 \left[\frac{\partial}{\partial T_{ab}} \boldsymbol{T}_{:,i}^\top \boldsymbol{C} \boldsymbol{T}_{:,j} \right] \boldsymbol{T}_{:,i}^\top \boldsymbol{C} \boldsymbol{T}_{:,j} \frac{1}{\overline{h}_i \overline{h}_j} + \left(\boldsymbol{T}_{:,i}^\top \boldsymbol{C} \boldsymbol{T}_{:,j} \right)^2 \left[\frac{\partial}{\partial T_{ab}} \frac{1}{\overline{h}_i \overline{h}_j} \right] \right\} \\
&= 2 \sum_j \langle \boldsymbol{C}_{a,:}, \boldsymbol{T}_{:,j} \rangle \boldsymbol{T}_{:,b}^\top \boldsymbol{C} \boldsymbol{T}_{:,j} \frac{1}{\overline{h}_b \overline{h}_j} + 2 \sum_i \langle \boldsymbol{C}_{:,a}, \boldsymbol{T}_{:,i} \rangle \boldsymbol{T}_{:,i}^\top \boldsymbol{C} \boldsymbol{T}_{:,b} \frac{1}{\overline{h}_i \overline{h}_b} \\
&- \sum_{ij} \left(\boldsymbol{T}_{:,i}^\top \boldsymbol{C} \boldsymbol{T}_{:,j} \right)^2 \left\{ \frac{\delta_{i=b} \overline{h}_j + \delta_{j=b} \overline{h}_i}{\overline{h}_i^2 \overline{h}_j^2} \right\} \\
&= 2 \sum_j \langle \boldsymbol{C}_{a,:}, \boldsymbol{T}_{:,j} \rangle \boldsymbol{T}_{:,b}^\top \boldsymbol{C} \boldsymbol{T}_{:,j} \frac{1}{\overline{h}_b \overline{h}_j} + 2 \sum_i \langle \boldsymbol{C}_{:,a}, \boldsymbol{T}_{:,i} \rangle \boldsymbol{T}_{:,i}^\top \boldsymbol{C} \boldsymbol{T}_{:,b} \frac{1}{\overline{h}_i \overline{h}_b} \\
&- \sum_j \left(\boldsymbol{T}_{:,b}^\top \boldsymbol{C} \boldsymbol{T}_{:,j} \right)^2 \frac{1}{\overline{h}_b^2 \overline{h}_j} - \sum_i \left(\boldsymbol{T}_{:,i}^\top \boldsymbol{C} \boldsymbol{T}_{:,b} \right)^2 \frac{1}{\overline{h}_i \overline{h}_b^2}
\end{aligned}
\tag{17}
$$

Consider now any admissible couplings $\boldsymbol{T}^{(1)} \in \mathring{\mathcal{U}}_n(\boldsymbol{h})$ and $\boldsymbol{T}^{(2)} \in \mathring{\mathcal{U}}_n(\boldsymbol{h})$, we want to prove that

$$
f(\boldsymbol{T}^{(1)}) \geq f(\boldsymbol{T}^{(2)}) + \langle \nabla_{\boldsymbol{T}} f(\boldsymbol{T}^{(2)}), \boldsymbol{T}^{(1)} - \boldsymbol{T}^{(2)} \rangle_F \tag{18}
$$

First observe that we have

$$
\begin{aligned}
&\langle \nabla_{\boldsymbol{T}} f(\boldsymbol{T}^{(2)}), \boldsymbol{T}^{(1)} - \boldsymbol{T}^{(2)} \rangle_F \\
&= \sum_{ab} (T_{ab}^{(1)} - T_{ab}^{(2)}) \left\{ 2 \sum_j \langle \boldsymbol{C}_{a,:}, \boldsymbol{T}_{:,j}^{(2)} \rangle \boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \frac{1}{\overline{h}_b^{(2)} \overline{h}_j^{(2)}} + 2 \sum_j \langle \boldsymbol{C}_{:,a}, \boldsymbol{T}_{:,j}^{(2)} \rangle \boldsymbol{T}_{:,j}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,b}^{(2)} \frac{1}{\overline{h}_j^{(2)} \overline{h}_b^{(2)}} \right\} \\
&- \sum_{ab} (T_{ab}^{(1)} - T_{ab}^{(2)}) \left\{ \sum_j \left(\boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \right)^2 \frac{1}{\overline{h}_b^{(2)2} \overline{h}_j^{(2)}} + \sum_j \left(\boldsymbol{T}_{:,j}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,b}^{(2)} \right)^2 \frac{1}{\overline{h}_j^{(2)} \overline{h}_b^{(2)2}} \right\} \\
&= 2 \sum_{bj} \left(\boldsymbol{T}_{:,b}^{(1)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \right) \left(\boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \right) \frac{1}{\overline{h}_b^{(2)} \overline{h}_j^{(2)}} + 2 \sum_{bj} \left(\boldsymbol{T}_{:,b}^{(1)\top} \boldsymbol{C}^\top \boldsymbol{T}_{:,j}^{(2)} \right) \left(\boldsymbol{T}_{:,j}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,b}^{(2)} \right) \frac{1}{\overline{h}_j^{(2)} \overline{h}_b^{(2)}} \\
&- 2 \sum_{bj} \left(\boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \right)^2 \frac{1}{\overline{h}_b^{(2)} \overline{h}_j^{(2)}} - 2 \sum_{bj} \left(\boldsymbol{T}_{:,j}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,b}^{(2)} \right)^2 \frac{1}{\overline{h}_j^{(2)} \overline{h}_b^{(2)}} \\
&- \sum_{bj} \left\{ \left(\boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \right)^2 + \left(\boldsymbol{T}_{:,j}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,b}^{(2)} \right)^2 \right\} \left\{ \frac{\overline{h}_b^{(1)}}{\overline{h}_b^{(2)2} \overline{h}_j^{(2)}} - \frac{1}{\overline{h}_b^{(2)} \overline{h}_j^{(2)}} \right\} \\
&= 2 \sum_{bj} \boldsymbol{T}_{:,b}^{(1)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \frac{1}{\overline{h}_b^{(2)} \overline{h}_j^{(2)}} + 2 \sum_{bj} \boldsymbol{T}_{:,b}^{(1)\top} \boldsymbol{C}^\top \boldsymbol{T}_{:,j}^{(2)} \boldsymbol{T}_{:,j}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,b}^{(2)} \frac{1}{\overline{h}_j^{(2)} \overline{h}_b^{(2)}} \\
&- \sum_{bj} \left\{ \left(\boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \right)^2 + \left(\boldsymbol{T}_{:,j}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,b}^{(2)} \right)^2 \right\} \frac{\overline{h}_b^{(1)}}{\overline{h}_b^{(2)2} \overline{h}_j^{(2)}} - 2 \mathcal{F}(\boldsymbol{T}^{(2)})
\end{aligned}
\tag{19}
$$

So the difference of both terms in equation 18 reads:

$$
\begin{aligned}
&f(\boldsymbol{T}^{(1)}) - f(\boldsymbol{T}^{(2)}) - \langle \nabla_{\boldsymbol{T}} f(\boldsymbol{T}^{(2)}), \boldsymbol{T}^{(1)} - \boldsymbol{T}^{(2)} \rangle_F \\
&= \sum_{bj} \left(\boldsymbol{T}_{:,b}^{(1)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(1)} \right)^2 \frac{1}{\overline{h}_b^{(1)} \overline{h}_j^{(1)}} + \sum_{bj} \left(\boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \right)^2 \frac{1}{\overline{h}_b^{(2)} \overline{h}_j^{(2)}} \\
&+ \sum_{bj} \left\{ \left(\boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \right)^2 + \left(\boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C}^\top \boldsymbol{T}_{:,j}^{(2)} \right)^2 \right\} \frac{\overline{h}_b^{(1)}}{\overline{h}_b^{(2)2} \overline{h}_j^{(2)}} \\
&- 2 \sum_{bj} \left(\boldsymbol{T}_{:,b}^{(1)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \right) \left(\boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \right) \frac{1}{\overline{h}_b^{(2)} \overline{h}_j^{(2)}} - 2 \sum_{bj} \left(\boldsymbol{T}_{:,b}^{(1)\top} \boldsymbol{C}^\top \boldsymbol{T}_{:,j}^{(2)} \right) \left(\boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C}^\top \boldsymbol{T}_{:,j}^{(2)} \right) \frac{1}{\overline{h}_j^{(2)} \overline{h}_b^{(2)}}
\end{aligned}
\tag{20}
$$

Then notice that for any (b, j), we have

$$
\begin{aligned}
&\left(\boldsymbol{T}_{:,b}^{(1)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \right)^2 \frac{1}{\overline{h}_b^{(1)} \overline{h}_j^{(2)}} + \left(\boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \right)^2 \frac{\overline{h}_b^{(1)}}{\overline{h}_b^{(2)2} \overline{h}_j^{(2)}} - 2 \left(\boldsymbol{T}_{:,b}^{(1)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \right) \left(\boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \right) \frac{1}{\overline{h}_b^{(2)} \overline{h}_j^{(2)}} \\
&= \left(\frac{\boldsymbol{T}_{:,b}^{(1)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)}}{\sqrt{\overline{h}_b^{(1)} \overline{h}_j^{(2)}}} - \boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \sqrt{\frac{\overline{h}_b^{(1)}}{\overline{h}_b^{(2)2} \overline{h}_j^{(2)}}} \right)^2 = (A_{bj} - B_{bj})^2
\end{aligned}
\tag{21}
$$

then similarly we have

$$
\begin{aligned}
&\left(\boldsymbol{T}_{:,b}^{(1)\top} \boldsymbol{C}^\top \boldsymbol{T}_{:,j}^{(2)} \right)^2 \frac{1}{\overline{h}_b^{(1)} \overline{h}_j^{(2)}} + \left(\boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C}^\top \boldsymbol{T}_{:,j}^{(2)} \right)^2 \frac{\overline{h}_b^{(1)}}{\overline{h}_b^{(2)2} \overline{h}_j^{(2)}} - 2 \left(\boldsymbol{T}_{:,b}^{(1)\top} \boldsymbol{C}^\top \boldsymbol{T}_{:,j}^{(2)} \right) \left(\boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C}^\top \boldsymbol{T}_{:,j}^{(2)} \right) \frac{1}{\overline{h}_b^{(2)} \overline{h}_j^{(2)}} \\
&= \left(\frac{\boldsymbol{T}_{:,b}^{(1)\top} \boldsymbol{C}^\top \boldsymbol{T}_{:,j}^{(2)}}{\sqrt{\overline{h}_b^{(1)} \overline{h}_j^{(2)}}} - \boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C}^\top \boldsymbol{T}_{:,j}^{(2)} \sqrt{\frac{\overline{h}_b^{(1)}}{\overline{h}_b^{(2)2} \overline{h}_j^{(2)}}} \right)^2 = (A'_{bj} - B'_{bj})^2
\end{aligned}
\tag{22}
$$

So we can express the equation 20 as

$$
\begin{aligned}
&f(\boldsymbol{T}^{(1)}) - f(\boldsymbol{T}^{(2)}) - \langle \nabla_{\boldsymbol{T}} \mathcal{F}^{GW}(\boldsymbol{T}^{(2)}), \boldsymbol{T}^{(1)} - \boldsymbol{T}^{(2)} \rangle_F \\
&= \sum_{bj} \left(\boldsymbol{T}_{:,b}^{(1)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(1)} \right)^2 \frac{1}{\overline{h}_b^{(1)} \overline{h}_j^{(1)}} + \sum_{bj} \left(\boldsymbol{T}_{:,b}^{(2)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \right)^2 \frac{1}{\overline{h}_b^{(2)} \overline{h}_j^{(2)}} \\
&- \sum_{bj} \left\{ \left(\boldsymbol{T}_{:,b}^{(1)\top} \boldsymbol{C} \boldsymbol{T}_{:,j}^{(2)} \right)^2 + \left(\boldsymbol{T}_{:,b}^{(1)\top} \boldsymbol{C}^\top \boldsymbol{T}_{:,j}^{(2)} \right)^2 \right\} \frac{1}{\overline{h}_b^{(1)} \overline{h}_j^{(2)}} + \sum_{bj} \left\{ (A_{bj} - B_{bj})^2 + (A'_{bj} - B'_{bj})^2 \right\}
\end{aligned}
\tag{23}
$$

Now let us suppose that the function $g_{\boldsymbol{C}}$ defined in equation 15 of Lemma 2 is convex on $\mathcal{U}(\boldsymbol{h}, \boldsymbol{h})$, hence including low-rank couplings of the form $\boldsymbol{U} = \boldsymbol{T} \boldsymbol{D}_{\overline{\boldsymbol{h}}}^{-1} \boldsymbol{T}^\top$. Given $\boldsymbol{U}^{(1)} = \boldsymbol{T}^{(1)} \boldsymbol{D}_{\overline{\boldsymbol{h}}^{(1)}}^{-1} \boldsymbol{T}^{(1)\top}$ and $\boldsymbol{U}^{(2)} = \boldsymbol{T}^{(2)} \boldsymbol{D}_{\overline{\boldsymbol{h}}^{(2)}}^{-1} \boldsymbol{T}^{(2)\top}$, the convexity of $g_{\boldsymbol{C}}$ implies that for any $\lambda \in [0, 1]$,

$$
\begin{aligned}
&g_{\boldsymbol{C}}(\lambda \boldsymbol{U}^{(1)} + (1 - \lambda) \boldsymbol{U}^{(2)}) \\
&= \operatorname{Tr} \left\{ \left(\lambda \boldsymbol{U}^{(1)} + (1 - \lambda) \boldsymbol{U}^{(2)} \right) \boldsymbol{C}^\top \left(\lambda \boldsymbol{U}^{(1)} + (1 - \lambda) \boldsymbol{U}^{(2)} \right) \boldsymbol{C} \right\} \\
&= \lambda^2 \operatorname{Tr} \left\{ \boldsymbol{U}^{(1)} \boldsymbol{C}^\top \boldsymbol{U}^{(1)} \boldsymbol{C} \right\} + (1 - \lambda)^2 \operatorname{Tr} \left\{ \boldsymbol{U}^{(2)} \boldsymbol{C}^\top \boldsymbol{U}^{(2)} \boldsymbol{C} \right\} + 2\lambda(1 - \lambda) \operatorname{Tr} \left\{ \boldsymbol{U}^{(1)} \boldsymbol{C}^\top \boldsymbol{U}^{(2)} \boldsymbol{C} \right\} \\
&\leq \lambda \operatorname{Tr} \left\{ \boldsymbol{U}^{(1)} \boldsymbol{C}^\top \boldsymbol{U}^{(1)} \boldsymbol{C} \right\} + (1 - \lambda) \operatorname{Tr} \left\{ \boldsymbol{U}^{(2)} \boldsymbol{C}^\top \boldsymbol{U}^{(2)} \boldsymbol{C} \right\}
\end{aligned}
\tag{24}
$$

implying e.g for $\lambda = \frac{1}{2}$, that

$$
\operatorname{Tr} \left\{ \boldsymbol{U}^{(1)} \boldsymbol{C}^\top \boldsymbol{U}^{(1)} \boldsymbol{C} \right\} \leq \frac{1}{2} \operatorname{Tr} \left\{ \boldsymbol{U}^{(1)} \boldsymbol{C}^\top \boldsymbol{U}^{(1)} \boldsymbol{C} \right\} + \frac{1}{2} \operatorname{Tr} \left\{ \boldsymbol{U}^{(2)} \boldsymbol{C}^\top \boldsymbol{U}^{(2)} \boldsymbol{C} \right\} \tag{25}
$$

where for instance

$$\begin{aligned}
\mathrm{Tr}\left\{\boldsymbol{U}^{(1)}\boldsymbol{C}^\top\boldsymbol{U}^{(2)}\boldsymbol{C}\right\} &= \mathrm{Tr}\left\{\boldsymbol{T}^{(1)}\boldsymbol{D}^{-1}_{\overline{\boldsymbol{h}}^{(1)}}\boldsymbol{T}^{(1)\top}\boldsymbol{C}^\top\boldsymbol{T}^{(2)}\boldsymbol{D}^{-1}_{\overline{\boldsymbol{h}}^{(2)}}\boldsymbol{T}^{(2)\top}\boldsymbol{C}\right\} \\
&= \mathrm{Tr}\left\{\boldsymbol{D}^{-1/2}_{\overline{\boldsymbol{h}}^{(2)}}\boldsymbol{T}^{(2)\top}\boldsymbol{C}^\top\boldsymbol{T}^{(1)}\boldsymbol{D}^{-1/2}_{\overline{\boldsymbol{h}}^{(1)}}\boldsymbol{D}^{-1/2}_{\overline{\boldsymbol{h}}^{(1)}}\boldsymbol{T}^{(1)\top}\boldsymbol{C}\boldsymbol{T}^{(2)}\boldsymbol{D}^{-1}_{\overline{\boldsymbol{h}}^{(2)}}\right\} \\
&= \|\boldsymbol{D}^{-1/2}_{\overline{\boldsymbol{h}}^{(1)}}\boldsymbol{T}^{(1)\top}\boldsymbol{C}\boldsymbol{T}^{(2)}\boldsymbol{D}^{-1}_{\overline{\boldsymbol{h}}^{(2)}}\|_F^2 \\
&= \sum_{ij}\left(\boldsymbol{T}^{(1)\top}_{:,i}\boldsymbol{C}\boldsymbol{T}^{(2)}_{:,j}\right)^2\frac{1}{\overline{h}^{(1)}_i\overline{h}^{(2)}_j} \\
&= \sum_{ij}\left(\boldsymbol{T}^{(2)\top}_{:,j}\boldsymbol{C}^\top\boldsymbol{T}^{(1)}_{:,i}\right)^2\frac{1}{\overline{h}^{(1)}_i\overline{h}^{(2)}_j}
\end{aligned} \tag{26}$$

The last equality holds as $\boldsymbol{T}^{(1)\top}_{:,i}\boldsymbol{C}\boldsymbol{T}^{(2)}_{:,j} = \boldsymbol{T}^{(2)\top}_{:,j}\boldsymbol{C}^\top\boldsymbol{T}^{(1)}_{:,i}$. Notice that using the same kind of relations we have

$$\sum_{bj}\frac{\left(\boldsymbol{T}^{(1)\top}_{:,b}\boldsymbol{C}\boldsymbol{T}^{(1)}_{:,j}\right)^2}{\overline{h}^{(1)}_b\overline{h}^{(1)}_j} = \sum_{bj}\frac{\left(\boldsymbol{T}^{(1)\top}_{:,j}\boldsymbol{C}^\top\boldsymbol{T}^{(1)}_{:,b}\right)^2}{\overline{h}^{(1)}_b\overline{h}^{(1)}_j} = \sum_{bj}\frac{\left(\boldsymbol{T}^{(1)\top}_{:,b}\boldsymbol{C}^\top\boldsymbol{T}^{(1)}_{:,j}\right)^2}{\overline{h}^{(1)}_b\overline{h}^{(1)}_j} \tag{27}$$

This way we can express the concavity inequality in equation 25 as follows

$$\sum_{bj}\frac{\left(\boldsymbol{T}^{(1)\top}_{:,b}\boldsymbol{C}\boldsymbol{T}^{(1)}_{:,j}\right)^2}{\overline{h}^{(1)}_b\overline{h}^{(1)}_j} + \sum_{bj}\frac{\left(\boldsymbol{T}^{(2)\top}_{:,b}\boldsymbol{C}\boldsymbol{T}^{(2)}_{:,j}\right)^2}{\overline{h}^{(2)}_b\overline{h}^{(2)}_j} - 2\sum_{bf}\frac{\left(\boldsymbol{T}^{(1)\top}_{:,b}\boldsymbol{C}\boldsymbol{T}^{(2)}_{:,j}\right)^2}{\overline{h}^{(1)}_b\overline{h}^{(2)}_j} \geq 0 \tag{28}$$

and symetrically using equation 27 as

$$\sum_{bj}\frac{\left(\boldsymbol{T}^{(1)\top}_{:,b}\boldsymbol{C}^\top\boldsymbol{T}^{(1)}_{:,j}\right)^2}{\overline{h}^{(1)}_b\overline{h}^{(1)}_j} + \sum_{bj}\frac{\left(\boldsymbol{T}^{(2)\top}_{:,b}\boldsymbol{C}^\top\boldsymbol{T}^{(2)}_{:,j}\right)^2}{\overline{h}^{(2)}_b\overline{h}^{(2)}_j} - 2\sum_{bf}\frac{\left(\boldsymbol{T}^{(1)\top}_{:,b}\boldsymbol{C}^\top\boldsymbol{T}^{(2)}_{:,j}\right)^2}{\overline{h}^{(1)}_b\overline{h}^{(2)}_j} \geq 0 \tag{29}$$

So we can conclude from equation 23, equation 28 and equation 29 that

$$\begin{aligned}
&f(\boldsymbol{T}^{(1)}) - f(\boldsymbol{T}^{(2)}) - \langle\nabla_{\boldsymbol{T}}f(\boldsymbol{T}^{(2)}), \boldsymbol{T}^{(1)} - \boldsymbol{T}^{(2)}\rangle_F \\
&\geq \sum_{bj}\left\{(A_{bj} - B_{bj})^2 + (A'_{bj} - B'_{bj})^2\right\} \\
&\geq 0
\end{aligned} \tag{30}$$

Hence it is enough to have $g_{\boldsymbol{C}}$ convex on $\mathcal{U}(\boldsymbol{h},\boldsymbol{h})$ to get f convex on $\overset{\circ}{\mathcal{U}}_n(\boldsymbol{h})$, and equivalently $\mathcal{F}$ concave on $\overset{\circ}{\mathcal{U}}_n(\boldsymbol{h})$. □

Proof of Theorem 1. Following Lemma 3, if $g_{\boldsymbol{C}}$ is convex on $\mathcal{U}(\boldsymbol{h},\boldsymbol{h})$ we know that $\mathcal{F}$ is concave on $\overset{\circ}{\mathcal{U}}_n(\boldsymbol{h})$. Moreover, we also proved in Lemma 1 that $\mathcal{F}$ is continuous on $\mathcal{U}_n(\boldsymbol{h})$.

Now let us consider any $\boldsymbol{E} \in \mathcal{U}_n(\boldsymbol{h})\backslash\overset{\circ}{\mathcal{U}}_n(\boldsymbol{h})$, *i.e* there exists at least one $i \in [\![n]\!]$, such that $\|\boldsymbol{E}_{:,i}\|_1 = 0$. Let any $\boldsymbol{T} \in \overset{\circ}{\mathcal{U}}_n(\boldsymbol{h})$, and any $\lambda \in [0,1]$. As $\mathcal{U}_n(\boldsymbol{h})$ is compact, we can define a sequence $\left\{\boldsymbol{V}^{(m)} = \frac{1}{m}\boldsymbol{T} + (1-\frac{1}{m})\boldsymbol{E}\right\}_{m\in\mathbb{N}}$ such that $\boldsymbol{V}^{(m)} \xrightarrow[m\to\infty]{} \boldsymbol{E}$. By construction, $\forall m$, $\boldsymbol{V}^{(m)} \in \overset{\circ}{\mathcal{U}}_n(\boldsymbol{h})$, as $\boldsymbol{V}^{(m)} \in \mathcal{U}_n(\boldsymbol{h})$ by convexity and $\forall i \in [\![n]\!]$, $\|\boldsymbol{V}^{(m)}_{:,i}\|_1 = \frac{1}{m}\|\boldsymbol{T}_{:,i}\|_1 + (1-\frac{1}{m})\|\boldsymbol{E}_{:,i}\|_1 > 0$. Then we have by concavity in $\overset{\circ}{\mathcal{U}}_n(\boldsymbol{h})$:

$$\mathcal{F}(\lambda\boldsymbol{T} + (1-\lambda)\boldsymbol{V}_m) \geq \lambda\mathcal{F}(\boldsymbol{T}) + (1-\lambda)\mathcal{F}(\boldsymbol{V}_m) \tag{31}$$

then by continuity of $\mathcal{F}$ on $\mathcal{U}_n(\boldsymbol{h})$, we have when $m \to \infty$,

$$\mathcal{F}(\lambda\boldsymbol{T} + (1-\lambda)\boldsymbol{E}) \geq \lambda\mathcal{F}(\boldsymbol{T}) + (1-\lambda)\mathcal{F}(E) \tag{32}$$

which holds for any $\boldsymbol{T} \in \overset{\circ}{\mathcal{U}}_n(\boldsymbol{h})$ and any $\lambda \in [0,1]$. Notice that the same reasoning can be done for $\boldsymbol{T} \in \mathcal{U}_n(\boldsymbol{h}) \backslash \overset{\circ}{\mathcal{U}}_n(\boldsymbol{h})$ by considering another analog sequence that converges to $\boldsymbol{T}$. So we might conclude that $\mathcal{F}$ is concave on $\mathcal{U}_n(\boldsymbol{h})$. Therefore problem srGW-bary2 is a concave problem over a polytope, hence admits extremities of $\mathcal{U}_n(\boldsymbol{h})$ as minimum, and so does srGW-bary1 thanks to Lemma 1. Notice that one can express extremities of $\mathcal{U}_n(\boldsymbol{h})$ as $\left\{\text{diag}(\boldsymbol{h})\boldsymbol{M} | \boldsymbol{M} \in \{0,1\}^{N \times n},\ \forall i \in [\![N]\!],\ \exists! j \in [\![n]\!], M_{ij} = 1\right\}$ [6, Theorem 1]. □

Extension to GW. Finally for the sake of completeness, we can follow an analog development for GW instead of srGW, i.e considering the barycenter distribution fixed to $\overline{\boldsymbol{h}} \in \Sigma_n^*$, leading to the following GW barycenter problem:

$$\min_{\overline{\boldsymbol{C}} \in \mathbb{R}^{n \times n}, \boldsymbol{T} \in \mathcal{U}(\boldsymbol{h}, \overline{\boldsymbol{h}})} \mathcal{E}^{GW}(\boldsymbol{C}, \overline{\boldsymbol{C}}, \boldsymbol{T}) \tag{GW-bary-1}$$

Usong the same notations than in Theorem 1, we can state the next result:

Corollary 1. *Let $\boldsymbol{C} \in \mathbb{R}^{N \times N}$ any bounded matrix, $\boldsymbol{h} \in \Sigma_N^*$ and $\overline{\boldsymbol{h}} \in \Sigma_n^*$. If the function $g_{\boldsymbol{C}}$ defined for any $\boldsymbol{U} \in \mathcal{U}(\boldsymbol{h}, \boldsymbol{h})$ as*

$$g_{\boldsymbol{C}}(\boldsymbol{U}) = \text{vec}(\boldsymbol{U})^\top \left(\boldsymbol{C} \otimes_K \boldsymbol{C}\right) \text{vec}(\boldsymbol{U}) \tag{33}$$

is <u>convex</u> on $\mathcal{U}(\boldsymbol{h}, \boldsymbol{h})$, then the following GW-bary-2 problem

$$\min_{\boldsymbol{T} \in \mathcal{U}(\boldsymbol{h}, \overline{\boldsymbol{h}})} \mathcal{E}^{GW}(\boldsymbol{C}, \widetilde{\boldsymbol{C}}(\boldsymbol{T}), \boldsymbol{T}) \tag{GW-bary-2}$$

is <u>concave</u>. Hence the GW-bary-1 problem admits extremities of $\mathcal{U}(\boldsymbol{h}, \overline{\boldsymbol{h}})$ as optimum.

Proof of Corollary 1. Assuming that $g_{\boldsymbol{C}}$ is convex on $\mathcal{U}(\boldsymbol{h}, \boldsymbol{h})$, implies that the srGW-bary2 problem is concave as the objective function $\boldsymbol{T} \to \mathcal{E}_2(\boldsymbol{C}, \widetilde{\boldsymbol{C}}(\boldsymbol{T}), \boldsymbol{T})$ is concave on $\mathcal{U}_n(\boldsymbol{h})$. Therefore this function is necessarily concave on $\mathcal{U}(\boldsymbol{h}, \overline{\boldsymbol{h}})$ which is a convex subset of $\mathcal{U}_n(\boldsymbol{h})$. So we can conclude that the GW-bary-2 problem is concave. □

C Additional Details for Methods and Experiments

C.1 Extension to the Fused Gromov-Wasserstein Framework

As mentioned in Section 3, we further propose to extend GW-DR to the Fused Gromov-Wasserstein framework in order to explicitly incorporate features $\boldsymbol{X}$. It reads as follows

$$\min_{\boldsymbol{Z} \in \mathbb{R}^{n \times d}, \overline{\boldsymbol{F}} \in \mathbb{R}^{n \times p}} \text{srFGW}_{\alpha, L}(\boldsymbol{C}_X, \boldsymbol{X}, \boldsymbol{h}_X, \boldsymbol{C}_Z, \overline{\boldsymbol{F}})\,. \tag{FGW-DR}$$

where $\text{srFGW}_{\alpha,L}$ relates to the semi-relaxed Fused Gromov-Wasserstein divergence parametrized by $\alpha \in [0,1]$ and the choice of inner-loss for L taken as L_2 or L_{KL}. Following notations in Section B, this divergence can be expressed as follows

$$\min_{\boldsymbol{T} \in \mathcal{U}_n(\boldsymbol{h}_X)} \alpha \sum_{ijkl} L([\boldsymbol{C}_X]_{ij}, [\boldsymbol{C}_Z]_{kl}) T_{ik} T_{jl} \ + \ (1-\alpha) \sum_{ijk} L\left(X_{ik}, \overline{F}_{jk}\right) T_{ij} \tag{34}$$

where $\mathcal{U}_n(\boldsymbol{h}_X) = \left\{\boldsymbol{T} \in \mathbb{R}_+^{N \times n} | \boldsymbol{T}\boldsymbol{1}_n = \boldsymbol{h}_X\right\}$. As such, srFGW aims at finding a (semi-relaxed) optimal coupling by minimizing an OT cost which is a trade-off of a Wasserstein cost between feature matrices and a GW cost between the similarity matrices. As such, FGW-DR comes down to regularizing the inner OT problem with a semi-relaxed Wasserstein barycenter problem. The latter essentially reduces to a concave problem, wherein the goal is to achieve K-means clustering on $\boldsymbol{X}$ [4]. We acknowledge that the authors do not address this problem from an optimization point of view. To this end, one can follow an analog scheme than in the proof of Theorem 1 in the Wasserstein setting. Similarly, the minimization w.r.t $\overline{\boldsymbol{F}}$ of the Wasserstein barycenter objective admits closed-form solutions given $\boldsymbol{T} \in \mathcal{U}_n(\boldsymbol{h}_X)$, denoted $\widetilde{F}(\boldsymbol{T})$ [41]. Problem FGW-DR then can be equivalently written as

$$\min_{\boldsymbol{Z} \in \mathbb{R}^{n \times d}, \boldsymbol{T} \in \mathcal{U}_n(\boldsymbol{h}_X)} \alpha \sum_{ijkl} L([\boldsymbol{C}_X]_{ij}, [\boldsymbol{C}_Z]_{kl}) T_{ik} T_{jl} \ + \ (1-\alpha) \sum_{ijk} L\left(X_{ik}, \widetilde{F}_{jk}(\boldsymbol{T})\right) T_{ij} \tag{35}$$

where the second term relates to a concave function w.r.t $\boldsymbol{T}$, hence acting as a concave regularization w.r.t to $\boldsymbol{T}$ of GW-DR.

C.2 Experiments

Datasets. We first provide details about the datasets used in Section 4.

Table 4: Dataset Details.

	Number of samples	Dimensionality	Number of classes
MNIST	10000	784	10
F-MNIST	10000	784	10
COIL	1440	16384	20

Grid search for fused GW. As the two terms appearing in fused GW [41] may have different scales, we have to test a quite wide spectrum of values. For Table 3, we use the following grid

$$\{0, 0.000001, 0.0003, 0.005, 0.1, 0.25, 0.5, 0.75, 0.9, 0.995, 0.9997, 0.999999, 1\} . \quad (36)$$

About the implementation of GWDR. To initialize the prototypes' position, we sample independent $\mathcal{N}(0, 1)$ coordinates. Similarly, we initialize the transport plans by sampling uniform random variables in $[0, 1]$ before normalizing such that the marginal constraint is satisfied.