GRAPH SIGNAL REPRESENTATION WITH WASSERSTEIN BARYCENTERS

Effrosyni Simou and Pascal Frossard

Ecole Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Laboratory (LTS4)

ABSTRACT

In many applications signals reside on the vertices of weighted graphs. Thus, there is the need to learn low dimensional representations for graph signals that will allow for data analysis and interpretation. Existing unsupervised dimensionality reduction methods for graph signals have focused on dictionary learning. In these works the graph is taken into consideration by imposing a structure or a parametrization on the dictionary and the signals are represented as linear combinations of the atoms in the dictionary. However, the assumption that graph signals can be represented using linear combinations of atoms is not always appropriate. In this paper we propose a novel representation framework based on non-linear and geometry-aware combinations of graph signals by leveraging the mathematical theory of Optimal Transport. We represent graph signals as Wasserstein barycenters and demonstrate through our experiments the potential of our proposed framework for low-dimensional graph signal representation.

***Index Terms*—** graph signal processing, dimensionality reduction, optimal transport, barycenters

1. INTRODUCTION

In many applications signals can be naturally represented on graphs. For instance, in social, transportation or brain networks. Graph Signal Processing [1] aims to process such signals by taking into account the underlying graph structure. The reason for this is that the connections of the graph reveal important information about the relationships between the nodes. Therefore, processing a signal defined on a graph of N nodes yields more intuitive results than processing its vectorized N-dimensional counterpart which ignores the underlying graph structure. Given a set of graph signals, it is interesting to identify the underlying processes. This problem is addressed by dimensionality reduction methods specifically developed for graph signals.

Unsupervised dimensionality reduction methods for graph signals have focused on dictionary learning methods where the graph is taken into consideration by imposing a structure on an overcomplete dictionary. The signals are then represented as linear combinations of a small number of atoms in the dictionary [2], [3], [4], [5], [6], [7]. In addition, in [8] a linear wavelet transform is proposed, where the wavelets are learned with an autoencoder in order to be adaptive to a given class of signals. In practice, however, graph signals cannot always be represented using linear combinations of features.

As a motivating example of the case where linear combinations of features may be ineffective, consider the case where the dataset may contain different instances of a diffusion process. If the dictionary contains atoms with two very distinct instances of the diffusion process, it is impossible to represent a signal at an intermediate time instance as a linear combination of those two atoms. Furthermore, increasing the number of atoms to include more instances of the diffusion process will not offer a good solution either, since the dictionary will then have a high coherence thus leading to decreased performance of the sparse coding step.

In this work we propose a representation framework with non-linear and geometry-aware combinations of graph signal features by leveraging the mathematical theory of Optimal Transport (OT) [9]. OT is a powerful mathematical theory that allows to define distances which exploit the geometry of the underlying space. It has thus found numerous applications, for example in image processing [10], [11], domain adaptation [12], traffic congestion control [13], minimum-cost network flow [14], [15] and graphics [16], [17]. In our present work, we unite Optimal Transport and Graph Signal Processing with the goal to further exploit the underlying structure in graph signal representation problems. By taking into account the graph through the shortest path distance between the nodes, we employ Wasserstein distances between graph signals and compute a Wasserstein barycenter [18] of a set of graph signals. We show that Wasserstein barycenters provide a geometry-aware representation of a set of graph signals. Through our experiments we demonstrate the superiority of Wasserstein barycenters compared to linear combinations for low-dimensional graph signal representation. To the best of our knowledge, our work is the first to exploit the benefits of the geometry awareness that Optimal Transport has to offer in order to enhance graph signal representation methods.

2. OPTIMAL TRANSPORT FRAMEWORK

Optimal Transport [19] is a mathematical theory that allows to compute distances between probability measures defined

on a space by leveraging the distance metric on that space.

The OT problem was formulated by Monge [20]. Given two probability distributions with densities $I_0, I_1 \geq 0$, $\int I_0 = \int I_1 = 1$ and their respective domains Ω_0 and Ω_1, the objective is to find a map $f : \Omega_0 \to \Omega_1$ that pushes I_0 to I_1 in the most efficient way with respect to a metric-dependent cost $c : \Omega_0 \times \Omega_1 \to \mathbb{R}^+$ and a mass preservation constraint MP. The Monge formulation of the OT problem is as follows:

$$M(I_0, I_1) = \inf_{f \in MP} \int_{\Omega_0} c(x, f(x)) I_0(x) dx \tag{1}$$

where MP is the mass-preservation constraint which for a subset $B \subset \Omega_1$ can be expressed as:

$$\int_{x: f(x) \in B} I_0(x) dx = \int_B I_1(y) dy \tag{2}$$

In some cases there is no transport map f that can rearrange I_0 to I_1. The relaxation proposed by Kantorovich [21] made the solution of the transportation problem tractable in that case by allowing for split of mass. The Kantorovich OT problem aims to find a transport plan γ, where $\gamma(x, y)$ describes the amount of mass moving from x to y, as follows:

$$K(I_0, I_1) = \min_{\gamma \in MP} \int_{\Omega_0 \times \Omega_1} c(x, y) d\gamma(x, y) \tag{3}$$

where the mass preservation constraint (MP) for $A \subseteq \Omega_0$ and $B \subseteq \Omega_1$ can now be expressed as:

$$\gamma(\Omega_0 \times B) = I_1(B) \text{ and } \gamma(A \times \Omega_1) = I_0(A) \tag{4}$$

In the case where mass can only be found at specific positions, probability vectors (histograms) can be considered and the metric-dependent cost is in the form of an $n \times m$ matrix C, where n is the dimensionality of the source histogram and m is the dimensionality of the target histogram. The Kantorovich OT problem then aims to find an optimal coupling Γ that will minimize the cost of the mass transportation with respect to the matrix C. The Kantorovich formulation of the OT problem for histograms is as follows:

$$K(\mathbf{I_0}, \mathbf{I_1}) = \min_{\Gamma \in MP} \langle C, \Gamma \rangle \tag{5}$$

where the mass preservation constraint (MP) now reads:

$$\Gamma \in \mathbb{R}_+^{n \times m} : \Gamma \mathbf{1}_m = \mathbf{I_0} \text{ and } \Gamma^\top \mathbf{1}_n = \mathbf{I_1} \tag{6}$$

where $\mathbf{I_0}$ is the n-dimensional source histogram and $\mathbf{I_1}$ is the m-dimensional target histogram.

If $n = m$, D is a distance on $[\![n]\!]$ and the cost matrix C in Eq. (5) can be expressed as $C = D^p$ for $p \geq 1$, then:

$$W_p(\mathbf{I}_0, \mathbf{I}_1) = \left(\min_{\Gamma \in MP} \langle D^p, \Gamma \rangle \right)^{1/p} \tag{7}$$

defines the p-Wasserstein distance on the simplex $\Sigma_n = \{\mathbf{a} \in \mathbb{R}_+^n : \sum_{i=1}^n \mathbf{a}_i = 1\}$.

3. WASSERSTEIN DISTANCE BETWEEN GRAPH SIGNALS

We now propose to use the OT framework to compare graph signals. We consider a connected, weighted and undirected graph $\mathcal{G}$ with no self-loops and non-negative weights. The set of N vertices of $\mathcal{G}$ is denoted as $\mathcal{V}$ and the set of its edges as $\mathcal{E}$. The connectivity pattern of the graph is expressed via its weight matrix $\mathcal{W}$. The element $\mathcal{W}(i, j)$ corresponds to the weight of the edge connecting vertices $i, j \in \mathcal{V}$.

We assume that signal intensities (mass) can only be found at the N vertices of the graph and therefore we consider graph signals to be histograms. Thus, all considered graph signals belong to the simplex Σ_N, where N is the number of vertices. Since we consider graph signals to be probability vectors, we only deal here with signals with non-negative values and normalize with the l_1 norm.

The geodesic distance of a graph captures the shortest path distance between vertices. If s_{ij} is the shortest path between two nodes i and j, the geodesic distance between them is defined as the sum of the weights of the edges in s_{ij}:

$$D_G(i, j) = \sum_{(k,l) \in s_{ij}} \mathcal{W}(k, l) \tag{8}$$

We now show that D_G satisfies the properties of a distance for the considered graph, so we can use it to define p-Wasserstein distances between graph signals with Eq. (7):

- **Non-negativity:** Since the graph is connected, there exists a path between any two pairs of nodes i, j. Therefore, $D_G(i, j) > 0, \forall i, j \in \mathcal{V}, i \neq j$. Furthermore, since there are no self-loops in the graph, it holds that $D_G(i, i) = 0$. Therefore D_G satisfies the condition of non-negativity.
- **Symmetry:** Since the considered graph $\mathcal{G}$ is undirected, the shortest path from node i to node j is the shortest path from j to i reversed. As a result $D_G(i, j) = D_G(j, i)$ and therefore, D_G is symmetric.
- **Identity of indiscernibles:** Since the considered graph is connected with no self-loops it holds that $D_G(i, j) = 0$ if and only if $i = j$. Therefore, D_G satisfies the identity of indiscernibles.
- **Triangle inequality:** Let $i, j, k \in \mathcal{V}$. If the shortest path from i to k passes through node j, then $D_G(i, k) = D_G(i, j) + D_G(j, k)$. If the shortest path from i to k does not pass through j, then $D_G(i, k) < D_G(i, j) + D_G(j, k)$. Therefore, D_G satisfies the triangle inequality $D_G(i, k) \leq D_G(i, j) + D_G(j, k)$.

The computational complexity of Wasserstein distances is $O(N^3 logN)$, where N is the number of nodes. Although this is not a particularly high complexity, it can become prohibitive in the case where we want to compute distances

between many graph signals. As this is the case for representation learning problems, in this work we use the entropy regularized Kantorovich problem proposed by M. Cuturi in [22]. This problem produces "smoothed" approximations of the transport plan Γ and can be solved efficiently in $O(N^2)$ with Sinkhorn projections. Furthermore, in our work we chose $p = 1$ because in that case the cost for the transportation of mass is equal to the geodesic distance. The W_1 distance between two δ functions located at vertices i and j is equal to $D_G(i, j)$ and, as a result, W_1 distances naturally quantify translations of graph kernels throughout the graph. Thus, we compute the distance between graph signals as:

$$W_1^{\alpha}(\mathbf{I_0}, \mathbf{I_1}) = \min_{\Gamma \in MP} \langle D_G, \Gamma \rangle + \alpha H(\Gamma) \tag{9}$$

where MP is the mass preservation constraint as in Eq. (6), $H(\Gamma) = \sum_{i,j} \Gamma_{ij} log(\Gamma_{ij} - 1)$ is the negative entropy of the transport plan Γ and α is the regularization parameter.

4. WASSERSTEIN BARYCENTERS FOR GRAPH SIGNAL REPRESENTATION

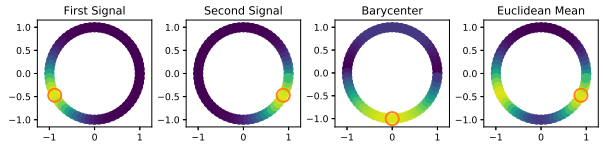
Equiped with the above definition of distances we now propose the use of Wasserstein barycenters [18] for the representation of graph signals. Given a set of signals, their barycenter can be thought of as their representative mean when using Wasserstein distances. Thus, given a set of M graph signal features $\{s_i\}_{i=1}^M$, we represent a graph signal x as the Wasserstein barycenter of $\{s_i\}_{i=1}^M$:

$$\hat{x} = \underset{u \in \Sigma_N}{\operatorname{argmin}} \sum_{i=1}^{M} \lambda_i W_1^{\alpha}(s_i, u) \tag{10}$$

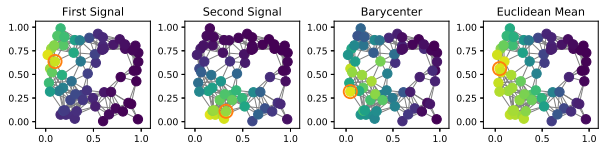
where W_1^{α} is the entropy regularized W_1 distance between graph signals as defined in Eq. (9) and λ_i is the weight corresponding to s_i. The intuition behind the use of Wasserstein barycenters for the representation of graph signals as in Eq. (10) is that the underlying graph structure is accounted for through the use of the W_1^{α} distances defined in Eq. (9). Therefore, we propose a geometry-aware combination of the composing features of the graph signal x.

We construct a regular ring graph and a sensor graph of $N = 64$ nodes. The sensor graph is created by placing randomly sensors in the xy-plane within the $[0, 1] \times [0, 1]$ square. The graph is created with an RBF kernel and hence the elements of the adjacency matrix are $W(i, j) = exp(-\frac{d_{ij}^2}{2\sigma^2})$, where d_{ij} is the distance between the nodes i and j.

We first consider a low-pass heat kernel ($\hat{g}_\tau(\lambda_l) = e^{-\tau\lambda_l}$) and we localize it at two distinct nodes of the graph. We then compute the entropy regularized barycenter of these two kernels as expressed in Eq. (10) for $\alpha = 0.001$ and $\lambda_1 = \lambda_2 = 0.5$. Because of the entropic regularization, the barycenter is expected to be more "spread out" than the initial signals, as explained in Section 3. The two signals, their barycenter and their Euclidean mean are shown in Fig. 1. It can be observed that the W_1^{α} barycenter is a localized kernel, similar to the initial heat kernels, albeit smoothed because of the entropic regularization. The Wasserstein distances between the graph kernels capture the geometry of the underlying domain and the barycenter resembles a heat kernel positioned between the two initial signals. The Wasserstein barycenter aims to find the most similar graph signal to the initial signals both in terms of the type of the graph signal and the position on the topology. On the contrary, Euclidean distances are completely ignorant of the underlying space and result to average representations of the graph signals that carry no geometric interpretation.



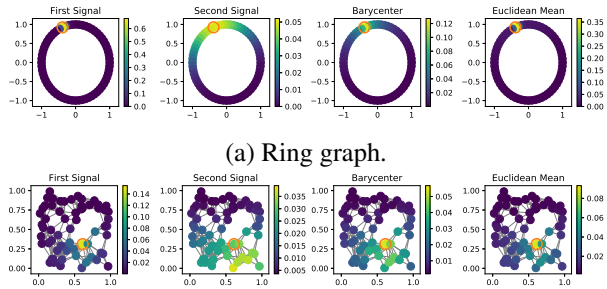
(a) Ring graph.



(b) Sensor graph.

Fig. 1. The localized heat kernels, their W_1^{α} barycenter and their Euclidean mean. The orange circle indicates the node where the kernel is localized for the two signals and the node with the highest signal value for the barycenter and the Euclidean mean.

We now consider two instances of a heat diffusion process starting at a node. The first is early on in the diffusion process (τ_1) and the second one is later (τ_2). We compute the entropy regularized W_1^{α} barycenter with $\alpha = 0.01$ and $\lambda_1 = \lambda_2 = 0.5$. The two graph signals, their W_1^{α} barycenter and their Euclidean mean are shown in Fig. 2. It can be observed that the W_1^{α} barycenter resembles to a heat diffusion process at an intermediate time instance τ ($\tau_1 \leq \tau \leq \tau_2$). This can be verified in the spectral domain, as can be seen in Fig. 3. For the first time instance τ_1 the signal is very localized and therefore its spectrum has components in almost all the frequencies. For the time instance τ_2 the process has diffused in most of the graph and therefore it is a smooth signal whose spectrum only contains information in the low frequencies. We would expect the diffusion process at an intermediate time step τ to have a spectrum that has more significant high frequency components than the second signal. This is the case for the W_1^{α} barycenter. In contrast, the Euclidean mean does not produce geometrically interpretable results, as can be seen in Fig. 2. It simply yields the average value of the signal values at each node and, as a result, completely ignores the "expansion" of the diffusion process through the graph.



(a) Ring graph.

(b) Sensor graph.

Fig. 2. The two instances of the heat diffusion process, their W_1^α barycenter and their Euclidean mean. The orange circle indicates the node where the heat diffusion process starts.

5. DIMENSIONALITY REDUCTION FOR GRAPH SIGNALS WITH WASSERSTEIN BARYCENTERS

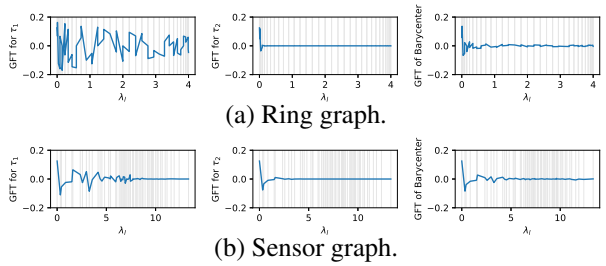
We now illustrate the potential of the Wasserstein barycenters introduced in Section 4 for dimensionality reduction of graph signals. For this purpose we adapt the Wasserstein Dictionary Learning (WDL) algorithm proposed in [11] for images in order to test whether it would yield meaningful features for the case of graph signals. The WDL problem is as follows:

Given a set of S signals $\{x_i\}_{i=1}^S$, learn the dictionary D and the weights Λ such that the difference between the signals and their representations as Wasserstein barycenters of the atoms of the dictionary is minimal according to a loss $\mathcal{L}$:

$$\min_{D,\Lambda} \mathcal{E}(D,\Lambda) := \sum_{i=1}^{S} \mathcal{L}(P(D,\Lambda_i), x_i) \tag{11}$$

where $P(D,\Lambda_i)$ is the representation of x_i as the barycenter of the atoms of D with weights Λ_i. By substituting $P(D,\Lambda_i)$ with our proposed Wasserstein barycenter representation for graph signals in Eq. (10) we can employ WDL [11] on graph signals.

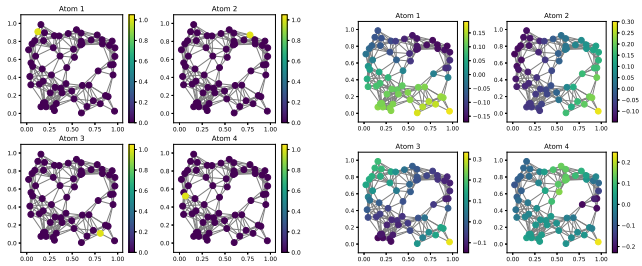
We consider the sensor graph of Section 4 and as training signals the graph signals obtained when we translate a localized heat kernel ($\tau = 5$) at each node of the graph. We solve the problem in Eq. (11) with $P(D,\Lambda_i)$ computed with Eq. (9)



(a) Ring graph.

(b) Sensor graph.

Fig. 3. Graph Fourier Transforms (GFT) of the two instances of the heat diffusion process and their W_1^α barycenter.



(a) WDL

(b) SVD

Fig. 4. Representations obtained with WDL and SVD.

and (10), $\mathcal{L} = L_2$ and learn a dictionary of $M = 4$ atoms. The atoms learned with WDL are shown in Fig. 4a. It can be observed that the atoms learned are very localized heat kernels (almost δ functions) located at four nodes positioned at the "edges" of the graph. This result is expected based on the intuition developed in Section 4. Since the entropy-regularized barycenter of two heat kernels resembles a smoother heat kernel positioned "in between", it was expected when solving the inverse problem of finding the graph kernels whose entropy regularized barycenters can represent heat kernels with $\tau = 5$ to recover sharper heat kernels localized at "extreme" positions on the graph. The atoms truly reveal the underlying process of the translation of a heat kernel along the graph.

When WDL [11] is applied on graph signals, the graph is accounted for through the Wasserstein barycenters, but no structure is imposed on the atoms of the dictionary. We thus chose a dimensionality reduction method that imposes no structure on the dictionary, but uses linear combinations instead of Wasserstein barycenters. Hence, we compare the representations obtained with WDL to those obtained with Singular Value Decomposition (SVD), shown in Fig. 4b. It can be observed that although the underlying process is simply the translation of a heat kernel, it is not captured by the representations of SVD. This is because in SVD the graph structure is not taken into consideration. We have therefore illustrated the benefit of geometry-aware combinations with Wasserstein barycenters compared to linear combinations that ignore the graph structure.

6. DISCUSSION AND FUTURE WORK

In this paper we have introduced the use of Optimal Transport in Graph Signal Processing with a view of further exploiting the underlying graph structure in representation learning problems for graph signals. We have developed a framework for computing distances between graph signals that take into account the underlying graph and we have proposed to use Wasserstein barycenters for dimensionality reduction of graph signals. In our future work we plan to incorporate the proposed framework in representation learning algorithms developed specifically for graph signals.

7. REFERENCES

[1] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, 2013.

[2] Xuan Zhang, Xiaowen Dong, and Pascal Frossard, "Learning of structured graph dictionaries," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 3373–3376.

[3] Dorina Thanou, David I Shuman, and Pascal Frossard, "Learning parametric dictionaries for signals on graphs," *IEEE Transactions on Signal Processing*, vol. 62, no. 15, pp. 3849–3862, 2014.

[4] Dorina Thanou and Pascal Frossard, "Multi-graph learning of spectral graph dictionaries," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[5] Yael Yankelevsky and Michael Elad, "Dual graph regularized dictionary learning," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 4, pp. 611–624, 2016.

[6] Yael Yankelevsky and Michael Elad, "Dictionary learning for high dimensional graph signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4669–4673.

[7] Yael Yankelevsky and Michael Elad, "Finding gems: Multi-scale dictionaries for high-dimensional graph signals," *arXiv preprint arXiv:1806.05356*, 2018.

[8] Raif Rustamov and Leonidas J Guibas, "Wavelets on graphs via deep learning," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 998–1006. 2013.

[9] Cédric Villani, *Optimal transport: old and new*, vol. 338, Springer Science & Business Media, 2008.

[10] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol, "Regularized discrete optimal transport," *SIAM Journal on Imaging Sciences*, vol. 7, no. 3, pp. 1853–1882, 2014.

[11] Morgan A. Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Maurice Ngolè Mboula, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck, "Wasserstein dictionary learning: Optimal transport-based unsupervised non-linear dictionary learning," *SIAM J. Imaging Sciences*, vol. 11, pp. 643–678, 2018.

[12] Nicolas Courty, Rémi Flamary, and Devis Tuia, "Domain adaptation with regularized optimal transport," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 274–289.

[13] Martin Beckmann, "A continuous model of transportation," *Econometrica: Journal of the Econometric Society*, pp. 643–660, 1952.

[14] Montacer Essid and Justin Solomon, "Quadratically regularized optimal transport on graphs," *SIAM Journal on Scientific Computing*, vol. 40, no. 4, pp. 1961–1986, 2018.

[15] Christian Léonard, "Lazy random walks and optimal transport on graphs," *The Annals of Probability*, vol. 44, no. 3, pp. 1864–1915, 2016.

[16] Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher, "Earth mover's distances on discrete surfaces," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, pp. 67, 2014.

[17] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas, "Convolutional wasserstein distances: Efficient optimal transportation on geometric domains," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, pp. 66, 2015.

[18] Martial Agueh and Guillaume Carlier, "Barycenters in the wasserstein space," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 2, pp. 904–924, 2011.

[19] Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde, "Optimal mass transport: Signal processing and machine-learning applications," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 43–59, 2017.

[20] Gaspard Monge, "Mémoire sur la théorie des déblais et des remblais," *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[21] Leonid Kantorovich, "On translation of mass," *Doklady Akademii nauk SSSR*, vol. 37, pp. 227–229, 1942.

[22] Marco Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems (NIPS)*, 2013, pp. 2292–2300.