Gromov-wasserstein averaging of kernel and distance matrices

Gromov-Wasserstein Averaging of Kernel and Distance Matrices

Gabriel Peyré GABRIEL.PEYRE@CEREMADE.DAUPHINE.FR
CNRS and Univ. Paris-Dauphine, Pl. du M. De Lattre De Tassigny, 75775 Paris 16, FRANCE

Marco Cuturi MCUTURI@I.KYOTO-U.AC.JP
Kyoto University, 36-1 Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, JAPAN

Justin Solomon JSOLOMON@MIT.EDU
Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Abstract

This paper presents a new technique for computing the barycenter of a set of distance or kernel matrices. These matrices, which define the interrelationships between points sampled from individual domains, are not required to have the same size or to be in row-by-row correspondence. We compare these matrices using the *softassign* criterion, which measures the minimum distortion induced by a probabilistic map from the rows of one similarity matrix to the rows of another; this criterion amounts to a regularized version of the Gromov-Wasserstein (GW) distance between metric-measure spaces. The barycenter is then defined as a Fréchet mean of the input matrices with respect to this criterion, minimizing a weighted sum of softassign values. We provide a fast iterative algorithm for the resulting nonconvex optimization problem, built upon state-of-the-art tools for regularized optimal transportation. We demonstrate its application to the computation of shape barycenters and to the prediction of energy levels from molecular configurations in quantum chemistry.

1. Introduction

Many classes of input data encountered in machine learning are best expressed using either pairwise distance matrices or kernels. For instance, the atoms making up proteins and other molecules have no canonical orientation and hence are only known up to rigid motion; so, the geometry of a molecule might be better described by the pairwise distances between atoms rather than by their absolute positions. Kernel and covariance matrices from different datasets exhibit similar structure, since they define data points by their roles relative to the rest of a dataset rather than in any absolute sense.

A major difficulty with this representation is that in many applications these matrices are not "registered" or "aligned," meaning that there is no explicit correspondence between their rows and columns. This is the case for shapes that undergo pose variation, articulation or deformation. Even worse, in some cases, similarities are not defined over the same ground space. For example, different molecules may have varying numbers of atoms. This inconsistency yields matrices of varying dimensions.

In this paper, we propose a theoretical and computational framework for summarizing collections of unaligned distance or kernel matrices. Building upon the notion of Gromov-Wasserstein (GW) distances between metric-measure spaces (Mémoli, 2007), we provide machinery for registering matrices of varying sizes *while* building interpolants and barycenters inheriting structure from the inputs. We also derive a fast regularized approximation scheme providing a practical and efficient way to recover these barycenters in practice.

1.1. Previous Work

Countless techniques leverage matrices of distances or kernels to describe a dataset. While distances and kernels have opposite ordering—a large distance indicates small similarity—unless it is necessary to distinguish we will use the term "similarity matrix" to refer to *any* matrix containing pairwise relationships.

Aligning similarity matrices. Comparison of similarity matrices is challenging due to a lack of alignment. Mathematically, they only define the *intrinsic* structure

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

of a dataset, which remains unchanged e.g. under rotation, translation, and isometric motion. More concretely, there is unlikely to be a canonical ordering of the rows or columns of similarity matrices, and removing or adding a few rows/columns may not significantly affect the structure. Automatic computation of a map aligning two similarity matrices is itself a challenging problem; our work is fairly unique in that we simultaneously compute such an alignment *while* solving an optimization problem over such matrices.

Matching rows of similarity matrices usually becomes a quadratic assignment problem (Loiola et al., 2007), which is NP-hard. For the specific case of aligning similarity matrices derived from pairwise distances in a metric space (e.g., geodesic distances along a surface), a well-known instance of this matching problem is computation of the Gromov-Hausdorff distance (Gromov, 2001), which was applied to matching in (Mémoli, 2007).

To relax the matching problem to a continuous—but still non-convex—optimization problem, the seminal work of Mémoli considers matching between metric-measure spaces, incorporating a probability distribution on the ground space that can account for some form of uncertainty (Mémoli, 2011). The geodesic properties of this "space of metric spaces" are considered in (Sturm, 2012). The recent work (Hendrikson, 2016) showcases the application of the corresponding Gromov-Wasserstein distance to clustering of biological and social networks. It is possible to further relax this formulation to a quadratic (Aflalo et al., 2015) or semidefinite (Kezurer et al., 2015) convex program, although this relaxation can fail if the input similarity matrices exhibit symmetries.

Rather than mixing the possible symmetric maps, we embrace the nonconvexity of the original problem and optimize quadratic matching objectives directly. After regularization, our matching subroutine resembles "softassign quadratic assignment" (Rangarajan et al., 1999; Gold & Rangarajan, 1996). We extend their iterations to general loss functions and to more advanced machine learning problems than nearest-neighbor search, in particular barycenter computation. We additionally leverage the fact that GW can compare arbitrary matrices, not just pairwise distances, making our machinery applicable to a broader range of tasks. The mapping component of our algorithm can be considered a generalized version of the method recently proposed in (Solomon et al., 2016).

Averaging similarity matrices. The design of an "average" or barycenter of similarity matrices is informed by two primary factors: A measure of discrepancy between similarity matrices and a definition of a mean. We will GW distances to define discrepancies, so our remaining task is to define what it means to compute a barycenter.

Even when similarity matrices are the same size and registered, it may not make sense to average them arithmetically. One natural way to define a mean or barycenter of inputs $(C_s)_s$ is to minimize the sum $\sum_s \lambda_s d^2(C, C_s)$ over all possible C, where d is some notion of distance. This generalized notion of a mean is known as the *Karcher* or *Fréchet* mean, which generalizes the notion of an average to a wide class of metric spaces (Nielsen & Bhatia, 2012).

Fréchet barycenters in Wasserstein space were proposed in (Agueh & Carlier, 2011). Several algorithms for their computation have been proposed including (Cuturi & Doucet, 2014; Benamou et al., 2015); these assume a fixed ground distance metric. Our method, however, finds more commonality with algorithms that compute Fréchet means in the space of covariance or metric matrices. (Dryden et al., 2009) considers several simple options for covariance matrices, mostly with fixed alignment; their "Procrustes" mean provides one rudimentary strategy for alignment. (Sra, 2011) defines a metric on registered positive definite kernel matrices that can be used for computing Fréchet means.

Barycenter computation is a building block for many learning methods. For instance, k-means clustering alternates between distances and barycenter computation. The nearest-centroid classifier provides a similar approach to classification (Manning et al., 2008).

Optimal transport (OT). OT (Villani, 2003) is a way to compare probability distributions (histograms in the finite-dimensional case) defined over either the same ground space or multiple pre-registered ground spaces. The means of comparison is a convex linear program optimizing for a matching that moves the mass from one distribution to the other with minimal cost. OT has been applied to machine learning problems including comparison of descriptors (Cuturi, 2013), k-means (Cuturi & Doucet, 2014), semi-supervised learning (Solomon et al., 2014), domain adaptation (Courty et al., 2014), designing loss functions (Frogner et al., 2015), low-rank approximation (Seguy & Cuturi, 2015), and dictionary learning (Rolet et al., 2016).

As highlighted above, GW-style matching extends OT to the case when the ground spaces are not pre-registered, yielding a non-convex quadratic program to compute the transport. The resulting transportation matrix can be understood is a soft registration from one domain to the other. Our algorithm solves a entropically-regularized version of this quadratic program by extending recent advances in the computation of OT (Cuturi, 2013; Benamou et al., 2015).

1.2. Contributions

Our first contribution is the definition of a new discrepancy between similarity matrices. It extends the "Gromov-Wasserstein" distance between metric-measure spaces to arbitrary matrices, using a generic loss functions to compare pairwise similarities and entropic regularization. It can be defined over different ground measured spaces (i.e. each point is equipped with a measure), which are not required to be registered a priori. Entropic regularization enables the design of a fast iterative algorithm to compute a stationary point of the non-convex energy defining the discrepancy.

Our second contribution is a new notion of the barycenter of a set of unregistered similarity matrices, defined as a Fréchet mean with respect to the GW discrepancy. We propose a block coordinate relaxation algorithm to compute a stationary point of the objective function defining our barycenter.

We showcase applications of our method to the computation of barycenters between shapes. We also exemplify how the GW discrepancy can be used to predict energy levels in quantum chemistry, where molecules are naturally represented using their Coulomb interaction matrices, a perfect fit for our unregistered dissimilarity matrix formalism.

The code to reproduce the results of this paper is available online.[1]

1.3. Notation

The simplex of histograms with N bins is $\Sigma_N \overset{\text{def.}}{=} \left\{p \in \mathbb{R}_N^+ \;;\; \sum_i p_i = 1\right\}$. The entropy of $T \in \mathbb{R}_+^{N\times N}$ is defined as $H(T) \overset{\text{def.}}{=} -\sum_{i,j=1}^N T_{i,j}(\log(T_{i,j}) - 1)$. The set of couplings between histograms $p \in \Sigma_{N_1}$ and $q \in \Sigma_{N_2}$ is

$$\mathcal{C}_{p,q} \overset{\text{def.}}{=} \left\{T \in (\mathbb{R}_+)^{N_1\times N_2} \;;\; T\mathbb{1}_{N_2} = p,\; T^\top \mathbb{1}_{N_1} = q\right\}.$$

Here, $\mathbb{1}_N \overset{\text{def.}}{=} (1,\ldots,1)^\top \in \mathbb{R}^N$. For any tensor $\mathcal{L} = (\mathcal{L}_{i,j,k,\ell})_{i,j,k,\ell}$ and matrix $(T_{i,j})_{i,j}$, we define the tensor-matrix multiplication as

$$\mathcal{L} \otimes T \overset{\text{def.}}{=} \left(\sum_{k,\ell} \mathcal{L}_{i,j,k,\ell} T_{k,\ell}\right)_{i,j}. \tag{1}$$

2. Gromov-Wasserstein Discrepancy

2.1. Entropic Optimal Transport

Optimal transport distances are useful to compare two histograms $(p,q) \in \Sigma_{N_1} \times \Sigma_{N_2}$ defined on the same metric space, or at least on spaces that have previously registered. Given some cost matrix $c \in \mathbb{R}_+^{N_1\times N_2}$, where $c_{i,j}$ represents the transportation cost between position indexed by i and j, we define the solution of entropically-regularized optimal transport between these two histograms as

$$\mathcal{T}(c,p,q) \overset{\text{def.}}{=} \operatorname*{argmin}_{T\in\mathcal{C}_{p,q}} \langle c, T\rangle - \varepsilon H(T), \tag{2}$$

which is a strictly convex optimization problem.

As shown in (Cuturi, 2013), the solution reads $\mathcal{T}(c,p,q) = \operatorname{diag}(a) K \operatorname{diag}(b)$ where $K \overset{\text{def.}}{=} e^{-\frac{c}{\varepsilon}} \in \mathbb{R}_+^{N_1\times N_2}$ is the so-called Gibbs kernel associated to c, and $(a,b) \in \mathbb{R}_+^{N_1}\times\mathbb{R}_+^{N_2}$ can be computed using Sinkhorn iterations

$$a \leftarrow \frac{p}{Kb} \quad \text{and} \quad b \leftarrow \frac{q}{K^\top a}, \tag{3}$$

where here $\frac{\cdot}{\cdot}$ denotes component-wise division.

2.2. Gromov-Wasserstein Discrepancy

Following the pioneering work of Mémoli, we consider input data expressed as metric-measure spaces (Mémoli, 2011). This corresponds to pairs of the form $(C,p) \in \mathbb{R}^{N\times N}\times\Sigma_N$, where N is an arbitrary integer (the number of elements in the underlying space). Here, C is a matrix representing either similarities or distances between these elements, and p is an histogram, which can account either for some uncertainty or relative importance between these elements. In case no prior information is known about a space, one can set $p = \frac{1}{N}\mathbb{1}_N$ to the uniform distribution. In our setting, since we target a wide range of machine-learning problems, we do not restrict the matrices C to be distance matrices, i.e., they are not necessarily positive and does not necessarily satisfy the triangle inequality.

We define the Gromov-Wasserstein discrepancy between two measured similarity matrices $(C,p) \in \mathbb{R}^{N_1\times N_1}\times\Sigma_{N_1}$ and $(\bar{C},q) \in \mathbb{R}^{N_2\times N_2}\times\Sigma_{N_2}$ as follows:

$$\mathrm{GW}(C,\bar{C},p,q) \overset{\text{def.}}{=} \min_{T\in\mathcal{C}_{p,q}} \mathcal{E}_{C,\bar{C}}(T) \tag{4}$$

$$\text{where} \quad \mathcal{E}_{C,\bar{C}}(T) \overset{\text{def.}}{=} \sum_{i,j,k,\ell} L(C_{i,k},\bar{C}_{j,\ell}) T_{i,j} T_{k,\ell}$$

The matrix T is a coupling between the two spaces on which the similarity matrices C and $\bar{C}$ are defined. Here L is some loss function to account for the misfit between the similarity matrices. Typical choices of loss include the quadratic loss $L(a,b) = L_2(a,b) \overset{\text{def.}}{=} \frac{1}{2}|a-b|^2$ and the Kullback-Leibler divergence $L(a,b) = \mathrm{KL}(a|b) \overset{\text{def.}}{=} a\log(a/b) - a + b$ (which is not symmetric). This definition (4) of GW extends slightly the one considered by (Mémoli, 2011), since we consider an arbitrary loss L (rather than just the L^2 squared loss). In the case

[1] https://github.com/gpeyre/2016-ICML-gromov-wasserstein

$L = L_2$, (Mémoli, 2011) proves that $\mathrm{GW}^{1/2}$ defines a distance on the space of metric measure spaces quotiented by measure-preserving isometries.

Introducing the 4-way tensor

$$\mathcal{L}(C, \bar{C}) \stackrel{\text{def.}}{=} (L(C_{i,k}, \bar{C}_{j,\ell}))_{i,j,k,\ell},$$

we notice that

$$\mathcal{E}_{C,\bar{C}}(T) = \langle \mathcal{L}(C, \bar{C}) \otimes T,\, T \rangle,$$

where $\otimes$ denotes the tensor-matrix multiplication (1).

The following proposition shows how to compute $\mathcal{L}(C, \bar{C}) \otimes T$ efficiently for a general class of loss functions:

Proposition 1. *If the loss can be written as*

$$L(a, b) = f_1(a) + f_2(b) - h_1(a)h_2(b) \tag{5}$$

for functions (f_1, f_2, h_1, h_2), then, for any $T \in \mathcal{C}_{p,q}$,

$$\mathcal{L}(C, \bar{C}) \otimes T = c_{C,\bar{C}} - h_1(C) T h_2(\bar{C})^\top. \tag{6}$$

where $c_{C,\bar{C}} \stackrel{\text{def.}}{=} f_1(C)p\mathbb{1}_{N_2}^\top + \mathbb{1}_{N_1} q^\top f_2(\bar{C})^\top$ is independent of T.

Proof. Under hypothesis (5), formula (1) shows that one has the decomposition $\mathcal{L}(C, \bar{C}) \otimes T = A + B + C$ where

$$A_{i,j} = \sum_k f_1(C_{i,k}) \sum_\ell T_{k,\ell} = (f_1(C)(T\mathbb{1}))_i$$
$$B_{i,j} = \sum_\ell f_2(\bar{C}_{j,\ell}) \sum_k T_{k,\ell} = (f_2(\bar{C})(T^\top \mathbb{1}))_j$$
$$C_{i,j} = \sum_k h_1(C_{i,k}) \sum_\ell h_2(\bar{C}_{j,\ell}) T_{k,\ell}$$

which is equal to $(h_1(C)(h_1(\bar{C})T^\top)^\top)_{i,j}$. □

Remark 1 (Computational complexity). Formula (6) shows that for this class of losses, one can compute $\mathcal{L}(C, \bar{C}) \otimes T$ efficiently in $O(N_1^2 N_2 + N_2^2 N_1)$ operations, using only matrix/matrix multiplications, instead of the $O(N_1^2 N_2^2)$ complexity of the naïve implementation of formula (1).

Remark 2 (Special cases). The square loss $L = L_2$ satisfies (5) for $f_1(a) = a^2$, $f_2(b) = b^2$, $h_1(a) = a$ and $h_2(b) = 2b$. The KL loss $L = \mathrm{KL}$ satisfies (5) for $f_1(a) = a\log(a) - a$, $f_2(b) = b$, $h_1(a) = a$ and $h_2(b) = \log(b)$.

2.3. Entropic Gromov-Wasserstein Discrepancy

We consider the following entropic approximation of the initial GW formulation (4)

$$\mathrm{GW}_\varepsilon(C, \bar{C}, p, q) \stackrel{\text{def.}}{=} \min_{T \in \mathcal{C}_{p,q}} \mathcal{E}_{C,\bar{C}}(T) - \varepsilon H(T). \tag{7}$$

This is a non-convex optimization problem. We propose to use projected gradient descent, where both the gradient step and the projection are computed according to the KL metric.

Iterations of this algorithm are given by

$$T \leftarrow \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{C}_{p,q}} \left(T \odot e^{-\tau(\nabla \mathcal{E}_{C,\bar{C}}(T) - \varepsilon \nabla H(T))} \right), \tag{8}$$

where $\tau > 0$ is a small enough step size, and the KL projector of any matrix K is

$$\mathrm{Proj}^{\mathrm{KL}}_{\mathcal{C}_{p,q}}(K) \stackrel{\text{def.}}{=} \operatorname*{argmin}_{T' \in \mathcal{C}_{p,q}} \mathrm{KL}(T'|K).$$

Proposition 2. *In the special case $\tau = 1/\varepsilon$, iteration (8) reads*

$$T \leftarrow \mathcal{T}(\mathcal{L}(C, \bar{C}) \otimes T, p, q). \tag{9}$$

Proof. As shown in (Benamou et al., 2015), the projection is nothing else than the solution to the regularized transport problem (2), hence

$$\mathrm{Proj}^{\mathrm{KL}}_{\mathcal{C}_{p,q}}(K) = \mathcal{T}(-\varepsilon \log(K), p, q).$$

One also has

$$\nabla \mathcal{E}_{C,\bar{C}}(T) - \varepsilon \nabla H(T) = \mathcal{L}(C, \bar{C}) \otimes T + \varepsilon \log(T).$$

Re-arranging the terms in (8), one obtains, in the special case $\tau\varepsilon = 1$, the desired formula. □

Iteration (9) defines a surprisingly simple algorithm, in which each update of T involves a Sinkhorn projection.

Remark 3 (Convergence). Using (Boţ et al., 2015) (see Theorem 12), iterations (8) are guaranteed to converge provided that τ is chosen small enough, $\tau < \tau_{\max}$. Unfortunately, in general, one does not have $\tau_{\max} > 1/\varepsilon$, so that the step size $\tau = 1/\varepsilon$ advocated by Proposition 2 is not covered by the theory. However, we found that using $\tau = 1/\varepsilon$ always leads to a converging sequence of T, and that this works well in practice. Note that when $L = L_2$, this recovers the "softassign quadratic assignement" algorithms defined in (Rangarajan et al., 1999; Gold & Rangarajan, 1996). The convergence proof (Rangarajan et al., 1999), however, only ensures convergence of the functional values (not of the iterates) and also only applies when the function being minimized in (7) is convex, which is not the case for arbitrary matrices (C, C_s).

3. Gromov-Wasserstein Barycenters

3.1. Gromov-Wasserstein Barycenters

We define Gromov-Wasserstein barycenters of measured similarity matrices $(C_s)_{s=1}^S$, where $C_s \in \mathbb{R}^{N_s \times N_s}$, using

a Fréchet mean formulation:

$$\min_{C\in\mathbb{R}^{N\times N}} \sum_s \lambda_s \,\mathrm{GW}_\varepsilon(C, C_s, p, p_s). \tag{10}$$

In this section, for simplicity, we assume that the base histograms $(p_s)_s$ and the histogram p associated to the barycenter are known and fixed. Note in particular that the size (N, N) of the targeted barycenter matrix should be fixed by the user. It is straightforward to extend the exposition as well as the optimization algorithm to the setting where p is unknown and included as an optimization variable.

This barycenter can be reformulated by re-introducing couplings

$$\min_{C,(T_s)_s} \sum_s \lambda_s \left(\mathcal{E}_{C,C_s}(T_s) - \varepsilon H(T_s)\right) \tag{11}$$

subject to the constraints that for all s, $T_s \in \mathcal{C}_{p,p_s} \subset \mathbb{R}_+^{N\times N_s}$.

Note that if L is convex with respect to its first variable, this problem is convex with respect to C but not with respect to $(T_s)_s$ (it is a quadratic but not necessarily positive problem with respect to the $(T_s)_s$).

We propose to minimize (11) using a block coordinate relaxation, i.e. iteratively minimizing with respect to the couplings $(T_s)_s$ and to the metric C.

Minimization with respect to $(T_s)_s$. The optimization problem (10) over $(T_s)_s$ alone decouples as S independent GW_ε optimizations

$$\forall\, s, \quad \min_{T_s\in\mathcal{C}_{p,p_s}} \mathcal{E}_{C,C_s}(T_s) - \varepsilon H(T_s).$$

A stationary point of this optimization problem can be computed using the optimization algorithm detailed in Section 2.3.

Minimization with respect to C. For given (T_s), the minimization with respect to C reads

$$\min_C \sum_s \lambda_s \langle \mathcal{L}(C, C_s) \otimes T,\, T\rangle. \tag{12}$$

The following proposition shows, for a large class of losses, how to compute the global minimizer in closed form.

Proposition 3. *If L satisfies (5) and f_1'/h_1' is invertible, then the solution to (12) reads*

$$C = \left(\frac{f_1'}{h_1'}\right)^{-1} \left(\frac{\sum_s \lambda_s T_s^\top h_2(C_s) T_s}{pp^\top}\right), \tag{13}$$

where we assume the normalization $\sum_s \lambda_s = 1$.

Proof. Using relation (6), the functional to be minimized reads

$$\sum_s \lambda_s \langle f_1(C) p\mathbb{1}^\top + \mathbb{1}p_s^\top f_2(C_s) - h_1(C) T_s h_2(C_s)^\top,\, T_s\rangle.$$

The first order optimality condition for this optimization problem thus reads $f_1'(C) \odot (pp^\top) = h_1'(C) \odot \sum_s \lambda_s T_s h_2(C_s) T_s^\top$, which gives the desired formula. □

The intuition underlying formula (13) is clear: Each $T_s^\top h_2(C_s) T_s$ is a "realigned" matrix where T_s acts as a fuzzy permutation (optimal transportation coupling) of both rows and columns of the distance matrix C_s. These realigned metric are then averaged, where the precise notion of "averaging" depends on the loss L.

Barycenters of PSD kernels using L^2 loss. For the square loss $L = L_2$, the update (13) becomes

$$C \leftarrow \frac{1}{pp^\top} \sum_s \lambda_s T_s^\top C_s T_s. \tag{14}$$

This formula highlights an important property of the method:

Proposition 4. *For $L = L_2$, if the $(C_s)_s$ are positive semidefinite (PSD) matrices, the iterates C produced by the algorithm are also PSD.*

Proof. Formula (14) shows that the update of C corresponds to a linear averaging of the matrices $(\mathrm{diag}(1/p) T_s^\top C_s T_s \,\mathrm{diag}(1/p))_s$, which are all PSD since the matrices $(C_s)_s$ are. □

This proposition implies, since the SDP cone is a closed set, that the output of our barycenter algorithm at convergence is an SDP kernel.

Barycenters of infinitely divisible kernels using KL loss. For the KL loss $L = \mathrm{KL}$, the update (13) similar, but with a geometric mean in place of an arithmetic mean

$$C \leftarrow \exp\left(\frac{1}{pp^\top} \sum_s \lambda_s T_s^\top \log(C_s) T_s\right). \tag{15}$$

The following proposition shows that this loss is particularly attractive for averaging infinitely divisible kernels. Recall that $C \in \mathbb{R}^{N\times N}$ is infinitely divisible if and only if $U = \log(C)$ is a conditionally positive semidefinite kernel, i.e. for all $x \in \mathbb{R}^N$ such that $\langle x, \mathbb{1}_N\rangle = 0$, then $\langle Ux,\, x\rangle \geqslant 0$.

Proposition 5. *For $L = \mathrm{KL}$, if the $(C_s)_s$ are infinitely divisible kernels, the iterates C produced by the algorithm are also infinitely divisible.*

Algorithm 1 Computation of GW_ε barycenters.

Input: $(C_s, p_s)_s, p$
Initialize C.
repeat
  // minimize over $(T_s)_s$
  for $s = 1$ **to** S **do**
    Initialize T_s.
    repeat
      // compute $c_s = \mathcal{L}(C, C_s) \otimes T_s$ *using* (6).
      $c_s \leftarrow f_1(C) + f_2(C_s)^\top - h_1(C) T_s h_2(C_s)^\top$
      // Sinkhorn iterations (3) *to compute* $\mathcal{T}(c_s, p, q)$
      Initialize $a \leftarrow \mathbb{1}$, set $K \leftarrow e^{-c_s/\varepsilon}$.
      repeat
        $a \leftarrow \frac{p}{Kb}, b \leftarrow \frac{q}{K^\top a}$.
      until convergence
      Update $T_s \leftarrow \mathrm{diag}(a) K \,\mathrm{diag}(b)$.
    until convergence
  end for
  // minimize over C *using* (13).
  $C \leftarrow \left(\frac{f_1'}{h_1'}\right)^{-1} \left(\frac{\sum_s \lambda_s T_s^\top h_2(C_s) T_s}{p p^\top}\right)$
until convergence

Proof. According to formula (15), one has to show that $U \overset{\text{def.}}{=} \sum_s \lambda_s \,\mathrm{diag}(1/p) T_s^\top U_s T_s \,\mathrm{diag}(1/p)$ is conditionally PSD prodived that the U_s matrices are. This is indeed the case, because, for any x such that $\langle x, \mathbb{1}_N \rangle = 0$, one has $\langle Ux, x \rangle = \sum_s \lambda_s \langle U_s x_s, x_s \rangle$ where $x_s \overset{\text{def.}}{=} T_s \frac{x}{p}$, and since $\langle x_s, \mathbb{1}_{N_s} \rangle = \langle \frac{x}{p}, T_s^\top \mathbb{1}_{N_s} \rangle = \langle \frac{x}{p}, p \rangle = 0$, one has $\langle U_s x_s, x_s \rangle \geqslant 0$ so that $\langle Ux, x \rangle \geqslant 0$, which proves the result. □

This proposition implies, since the cone of infinitely divisible kernels is a closed set, that the output of our barycenter algorithm at convergence is infinitely divisible.

A important example of infinitely divisible kernels is given by the form $C = e^{-\frac{D^2}{\sigma^2}}$, where $D_{i,j} = \|x_i - x_j\|$ is the Euclidean distance matrix between points $(x_i)_i$, and $\sigma > 0$ is a bandwidth parameter that controls the "locality" of the resulting kernel. One verifies that $\mathrm{KL}(e^{-\frac{D^2}{\sigma^2}} | e^{-\frac{\bar{D}^2}{\sigma^2}}) = \frac{1}{2\sigma^4} \|D^2 - \bar{D}^2\|^2 + o(1/\sigma^4)$, so that in the regime of large bandwidth $\sigma \to +\infty$, using the KL loss in conjunction with this class of kernels becomes equivalent to using the L^2 loss between squared Euclidean matrices. A similar result in fact holds for any smooth Bregman divergence loss.

Pseudocode. Algorithm 1 details the steps of the optimization technique. Note that it makes use of three nested iterations: (i) blockwise coordinate descent on $(T_s)_s$ and C, (ii) projected gradient descent on each T_s, and (iii) Sinkhorn iterations to compute the projection.

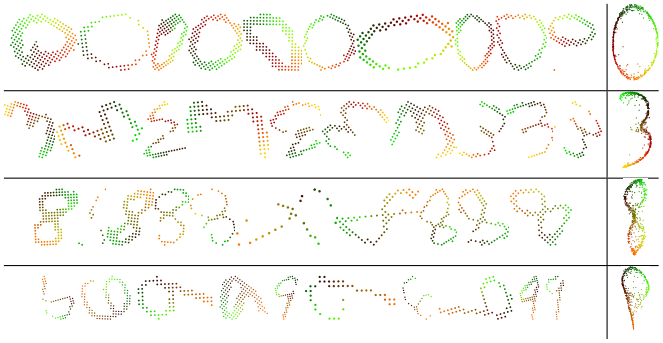


Figure 1. Barycenters of point clouds from MNIST digits; sample input data clouds are shown on the left, and an MDS embedding of the barycenter distance matrix is shown on the right.

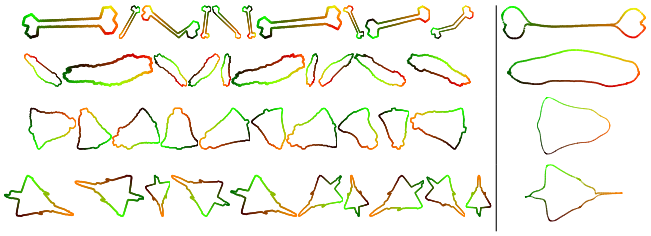


Figure 2. Barycenter example for shape data from (Thakoor et al., 2007).

4. Experiments

4.1. Point Clouds

Embedded barycenters. Figure 1 provides an example illustrating the behavior of our GW barycenter approximation. In this experiment, we extract 500 point clouds of handwritten digits from the dataset (LeCun et al., 1998), rotated arbitrarily in the plane. We represent each digit as a symmetric Euclidean distance matrix and optimize for a 500×500 barycenter using Algorithm 1 (uniform weights, $\varepsilon = 1 \times 10^{-3}$); notice that most of the input point clouds consist of fewer than 500 points. We then visualize the barycenter matrix as a point cloud in the plane using multi-dimensional scaling (MDS). Each digit is colored by transferring RGB values from the barycenter point cloud using the computed map T.

This experiment illustrates a few properties of our algorithm. Most prominently, note that the input digits are *rotated arbitrarily*, which—unlike tools based on classical optimal transportation—does not affect the computation of the barycenter. Also, to avoid bias in this experiment we initialize C as a random matrix, and yet our nonconvex optimization algorithm still reaches a meaningful barycenter.

Figure 2 illustrates a similar experiment on the 2D shape data from (Thakoor et al., 2007). This second experiment illustrates some additional properties of our barycenters. Interestingly, even though we do not impose curve topol-

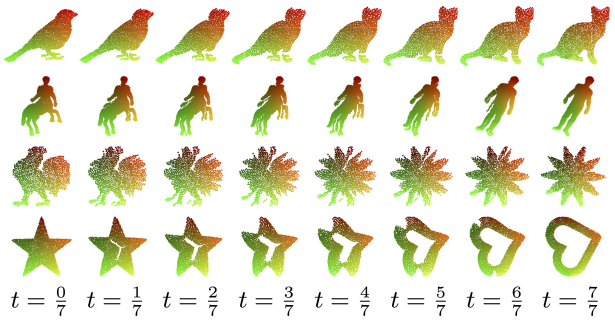
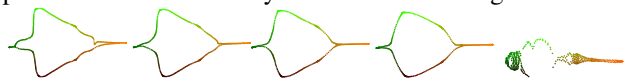


Figure 3. Progressive interpolation $(\bar{x}_t)_{t=0}^1$ between two shapes (x_0, x_1) using GW barycenters.

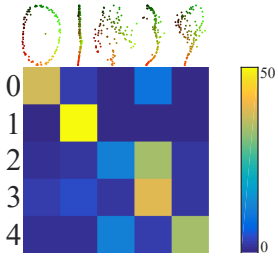
ogy on the barycenter, the final MDS embedding has one-dimensional structure. The barycenter is also resilient to noise in the computed map from individual examples, e.g. the bones in the first row.

Regularization. Figure bellow illustrates the dependence of our barycenter on the regularizer ε.



It shows the barycenter from a test in Figure 2 for increasing values of ε. When ε is small, the embedded barycenter most closely resembles the input data. As ε increases, the barycenters become smoother, with the advantage of higher resilience to noise and fewer iterations needed for optimization.

Clustering. The figure on the right illustrates the incorporation of our barycenter technique into a larger machine learning pipeline. It shows results of a k-means unsupervised clustering of 250 handwritten characters.



In this example, we construct a dataset with 50 point clouds from each of digits 0 to 4 from the example in Figure 1. We then cluster the point clouds by representing them as pairwise distance matrices and applying the k-means algorithm ($k = 5$), with k-means++ initialization (Arthur & Vassilvitskii, 2007). We underscore that each handwritten character is represented using an unaligned point cloud with varying numbers of points.

Each row in the figure corresponds to a different character, and each column corresponds to a different cluster center (MDS embedding shown above the table). Color is proportional to the number of data points assigned to each cluster. Even without supervision, this rudimentary technique for clustering finds meaningful clusters of handwritten digits from the dataset. All the "1" digits are clustered correctly. Digits 2 and 4 are mixed considerably in the computed clustering, reflecting the wide variety of handwritten styles for these two digits.

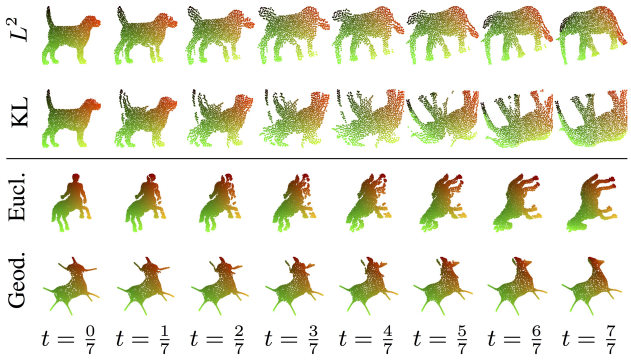


Figure 4. Upper part: comparison between interpolation using L^2 loss and KL loss. Lower part: comparison between interpolation using pairwise Euclidean and inner geodesic distances.

Shape interpolation. Figure 3 shows an application of the computation of GW barycenters between $S = 2$ input matrices two perform shape interpolation. The input matrices $(C_1, C_2) \in (\mathbb{R}^{N\times N})^2$ are Euclidean distance matrices of two input planar clouds $(x_{s,i})_i$ for $s \in \{1, 2\}$ uniformly sampled inside a shape, so that we use $p_1 = p_2 \frac{1}{N} \mathbb{1}_N$. This means that $C_{s,i,j} \overset{\text{def.}}{=} \|x_{s,i} - x_{s,j}\|$ for $s \in \{1, 2\}$ and $i, j = 1, \ldots, N$. The interpolation is achieved by first computing a barycenter matrix C_t minimizing (10) with our optimization algorithm, for $\lambda = (t, 1 - t)$, $t \in [0, 1]$, using the Euclidean loss $L = L_2$. A barycentric point cloud $\bar{x}_t = (\bar{x}_{t,i})_{i=1}^N$ is then reconstructed using SMACOF multidimensional scaling (MDS), which amounts to computing a local minimizer of $\sum_{i,j} |C_{t,i,j} - \|\bar{x}_{t,i} - \bar{x}_{t,j}\|^2$ with respect to $(\bar{x}_{t,i})_i$. This local minimizer $\bar{x}_t$ is computed by a gradient descent scheme, initialized with x_1.

L^2 vs. KL losses. Figure 4, upper part, shows a comparison of the same interpolation as those computed in Figure 3, computed with two different losses. Row #1 shows barycenters of the distance matrices $C_{s,i,j} = \|x_{s,i} - x_{s,j}\|$, for $s \in \{1, 2\}$, using the squared Euclidean loss $L = L_2$ as in the previous paragraph. The display is obtained by applying SMACOF MDS to the barycenter matrix C_t. Row #2 shows barycenter of the infinitely divisible kernels $C_{s,i,j} = \exp(-\|x_{s,i} - x_{s,j}\|^2/\sigma^2)$, for $s = \{1, 2\}$, using the Kullback-Leibler loss $L =$KL, and a bandwidth $\sigma = 1$ (for point clouds in the unit square). The display is then obtained by applying SMACOF MDS to $\sqrt{-\sigma \log(C_t)}$. The KL interpolation produces less regular interpolation because it imposes only local constraints at the scale of the bandwidth.

Euclidean vs. Geodesic distance matrices. Euclidean pairwise distance, though simple to compute and to manipulate, fail to capture the intrinsic geometry of shapes, and thus leads to physically unrealistic shape interpolation, that ignore in particular articulation. Figure 4, lower part, shows how this issue can be in large part alleviated by replacing the extrinsic Euclidean distance $C_{s,i,j} = \|x_{s,i} - x_{s,j}\|$ for $s \in \{1, 2\}$ by the inner geodesic distance (i.e. the length of the shortest path joining $x_{s,i}$ to $x_{s,j}$ while staying inside the shape). Note that, because the corresponding distance is not Euclidean anymore, the SMACOF MDS only produces an approximate embedding, the so-called bending invariant (Elad & Kimmel, 2003) used to perform isometric-invariant shape recognition.

4.2. Quantum chemistry

To demonstrate the interest of the Gromov-Wasserstein discrepancy, we consider its application to a regression problem on molecules. Several recent works (Rupp et al., 2012; Hansen et al., 2013) have proposed to predict atomization energies for molecules using descriptors of labeled molecules, rather than estimating them through expensive numerical simulations. (Rupp et al., 2012) proposed in particular to represent each molecule in the *qm7* dataset of organic 7165 molecules through its so-called "Coulomb matrix," which we describe next.

Coulomb matrices. For a molecule s with N_s atoms, each described by a relative location $r_i \in \mathbb{R}^3$ in space and a nuclear charge Z_i, Rupp et al. proposed to form the $N_s \times N_s$ matrix C_s with off-diagonal terms $Z_i Z_j / \|r_i - r_j\|$ for $1 \leqslant i \neq j \leqslant N_s$, and diagonal terms $\frac{1}{2} Z_i^{2.4}$. Although this plays no major role in our analysis, we found that these 7165 Coulomb matrices are all infinitely divisible positive definite kernel matrices. Rupp et al. argue that the Coulomb matrix is a convenient descriptor for molecules in the sense that it is invariant with respect to translations and rotations by construction. It has, however, an important weakness: C_s has an arbitrary ordering of the N_s atoms in s. As we explain next, this was addressed by Rupp et al. by creating randomly permuted copies of these matrices.

Previous work. Because their approach requires manipulating matrices of the same size, Rupp et al. pad all $N_s \times N_s$ Coulomb matrices C_s with zeros to obtain 23×23 matrices (23 is the maximal number of atoms in a molecule in the *qm7* database). To cope with the ordering problem, they also propose, as a representation for each molecule, to form several randomly permuted copies of C_s for each molecule s (between 8 and 1000 in (Hansen et al., 2013)). (Hansen et al., 2013, Table 3) provide out-of-sample mean absolute error (MAE) predictions for several techniques using 5 fold cross-validation. We report some of them in Table 4.2.

Algorithm	MAE	RMSE
k-nearest neighbors	71.54	95.97
Linear regression	20.72	27.22
Gaussian kernel ridge regression	8.57	12.26
Laplacian kernel ridge regression (8)	3.07	4.84
Multilayer Neural Network (1000)	3.51	5.96
GW 3-nearest neighbors	10.83	29.27

Table 1. Mean-Absolute and Root Mean Squared errors for the atomization energy prediction in the qm7 database of 7165 molecules. All results quoted from (Hansen et al., 2013, Table 3) except ours. Number in between parenthesis stand for random copies used in the algorithm. Although the GW distance is not directly competitive performance-wise, this result shows that an acceptable performance on this task can be recovered exclusively using simple metric tools.

Our approach. We propose to use the original $N_s \times N_s$ Coulomb matrices C_s directly as inputs for our GW distances. Namely, we compute a discrepancy between two molecules s, s' that is exactly $\mathrm{GW}(C_s, C_{s'}, \mathbb{1}_{N_s}/N_s, \mathbb{1}_{N_{s'}}/N_{s'})$, using $L = L_2$ and a regularization strength ε tuned heuristically on a subset of 500 points of the database. Using a 3-nearest neighbor regression approach we obtain a MAE of 10.83 which, although not directly competitive with the best approach recorded in (Hansen et al., 2013), it remarkable in the sense that it is obtained with an extremely simple classifier. Our result contradicts thus the observation made by Hansen et al., p.3414, that *This* [poor performance of k-nn] *indicates that there are meaningful linear relations between physical quantities in the system and that it is insufficient to simply look up the most similar molecules (as k- nearest neighbors does). The k-nearest neighbors approach fails to create a smooth mapping to the energies.*

5. Conclusion

The Gromov-Wasserstein discrepancy measure is relevant to countless problems in machine learning for which the primary role of a given data point is defined relative to other members of the dataset. After introducing an entropic regularizer to the problem, we in particular find that GW is not only theoretically attractive but also a practical tool with an efficient and easily-implemented non-convex optimization scheme. Our approach highlights the fact that using unregistered similarity matrices as features in machine learning problems is fruitful in many contexts. GW provides a natural framework for manipulation of these matrices, as we demonstrate on classification, clustering, and regression tasks for shapes and molecules. We suspect this elegant and efficient solution to extend to many other problem instances endowed with natural geometric structure.

Acknowledgements The work of G. Peyré has been supported by the European Research Council (ERC project SIGMA-Vision). J. Solomon acknowledges the support of the NSF Mathematical Sciences Postdoctoral Research Fellowship (award number 1502435).

References

Aflalo, Yonathan, Bronstein, Alexander, and Kimmel, Ron. On convex relaxation of graph isomorphism. *Proc. National Academy of Sci.*, 112(10):2942–2947, 2015.

Agueh, Martial and Carlier, Guillaume. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

Arthur, David and Vassilvitskii, Sergei. K-means++: The advantages of careful seeding. In *Proc. SODA*, pp. 1027–1035, 2007.

Benamou, Jean-David, Carlier, Guillaume, Cuturi, Marco, Nenna, Luca, and Peyré, Gabriel. Iterative Bregman projections for regularized transportation problems. *SIAM J. Sci. Comp.*, 37 (2):A1111–A1138, 2015.

Boţ, Radu Ioan, Csetnek, Ernö Robert, and László, Szilárd Csaba. An inertial forward–backward algorithm for the minimization of the sum of two nonconvex functions. *EURO J. Comp. Optim.*, pp. 1–23, 2015.

Courty, Nicolas, Flamary, Rémi, and Tuia, Devis. Domain adaptation with regularized optimal transport. In *Machine Learning and Knowledge Discovery in Databases*, pp. 274–289. 2014.

Cuturi, Marco. Sinkhorn distances: Lightspeed computation of optimal transportation. In *Proc. NIPS*, volume 26, pp. 2292–2300. 2013.

Cuturi, Marco and Doucet, Arnaud. Fast computation of Wasserstein barycenters. In *Proc. ICML*, volume 32, 2014.

Dryden, Ian L, Koloydenko, Alexey, and Zhou, Diwei. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, pp. 1102–1123, 2009.

Elad, Asi and Kimmel, Ron. On bending invariant signatures for surfaces. *IEEE Tr. on PAMI*, 25(10):1285–1295, 2003.

Frogner, Charlie, Zhang, Chiyuan, Mobahi, Hossein, Araya, Mauricio, and Poggio, Tomaso. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2044–2052. 2015.

Gold, Steven and Rangarajan, Anand. A graduated assignment algorithm for graph matching. *PAMI*, 18(4):377–388, April 1996.

Gromov, Mikhail. *Metric Structures for Riemannian and Non-Riemannian Spaces*. Progress in Math. Birkhäuser, 2001.

Hansen, Katja, Montavon, Grégoire, Biegler, Franziska, Fazli, Siamac, Rupp, Matthias, Scheffler, Matthias, Von Lilienfeld, O Anatole, Tkatchenko, Alexandre, and Mller, Klaus-Robert. Assessment and validation of machine learning methods for predicting molecular atomization energies. *Journal of Chemical Theory and Computation*, 9(8):3404–3419, 2013.

Hendrikson, Reigo. Using Gromov-Wasserstein distance to explore sets of networks. In *University of Tartu, Master Thesis*, 2016.

Kezurer, Itay, Kovalsky, Shahar Z., Basri, Ronen, and Lipman, Yaron. Tight relaxation of quadratic matching. *CGF*, 2015.

LeCun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Loiola, Eliane Maria, de Abreu, Nair Maria Maia, Boaventura-Netto, Paulo Oswaldo, Hahn, Peter, and Querido, Tania. A survey for the quadratic assignment problem. *European J. Operational Research*, 176(2):657–690, 2007.

Manning, Christopher D., Raghavan, Prabhakar, and Schütze, Hinrich. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Mémoli, Facundo. On the use of Gromov–Hausdorff distances for shape comparison. In *Symposium on Point Based Graphics*, pp. 81–90. 2007.

Mémoli, Facundo. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Comp. Math.*, 11 (4):417–487, 2011.

Nielsen, Frank and Bhatia, Rajendra. *Matrix Information Geometry*. Springer, 2012.

Rangarajan, Anand, Yuille, Alan, and Mjolsness, Eric. Convergence properties of the softassign quadratic assignment algorithm. *Neural Comput.*, 11(6):1455–1474, August 1999.

Rolet, Antoine, Cuturi, Marco, and Peyré, Gabriel. Fast dictionary learning with a smoothed Wasserstein loss. In *Proc. AISTATS'16*, 2016.

Rupp, Matthias, Tkatchenko, Alexandre, Müller, Klaus-Robert, and Von Lilienfeld, O Anatole. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5):058301, 2012.

Seguy, Vivien and Cuturi, Marco. Principal geodesic analysis for probability measures under the optimal transport metric. In *Advances in Neural Information Processing Systems*, pp. 3294–3302, 2015.

Solomon, Justin, Rustamov, Raif, Guibas, Leonidas, and Butscher, Adrian. Earth mover's distances on discrete surfaces. *ACM Trans. Graph.*, 33(4):67:1–67:12, July 2014.

Solomon, Justin, Peyré, Gabriel, Kim, Vladimir, and Sra, Suvrit. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (TOG)*, 35(4), 2016.

Sra, Suvrit. Positive definite matrices and the S-divergence. *arXiv preprint arXiv:1110.1773*, 2011.

Sturm, Karl-Theodor. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. Preprint 1208.0434, arXiv, 2012.

Thakoor, Ninad, Gao, Jean, and Jung, Sungyong. Hidden Markov model-based weighted likelihood discriminant for 2-D shape classification. *Trans. Image Proc.*, 16(11):2707–2719, 2007.

Villani, Cedric. *Topics in Optimal Transportation*. Graduate studies in Math. AMS, 2003.