Noname manuscript No.
(will be inserted by the editor)

Sliced and Radon Wasserstein Barycenters of Measures

Nicolas Bonneel · Julien Rabin · Gabriel Peyré · Hanspeter Pfister

the date of receipt and acceptance should be inserted later

Abstract This article details two approaches to compute barycenters of measures using 1-D Wasserstein distances along radial projections of the input measures. The first method makes use of the Radon transform of the measures, and the second is the solution of a convex optimization problem over the space of measures. We show several properties of these barycenters and explain their relationship. We show numerical approximation schemes based on a discrete Radon transform and on the resolution of a non-convex optimization problem. We explore the respective merits and drawbacks of each approach on applications to two image processing problems: color transfer and texture mixing.

Keywords Optimal transport · Radon transform · Wasserstein distance · Barycenter of measures

1 Introduction

The mass transportation problem corresponds to the computation of an optimal warping to map (i.e. push-forward) a given input probability measure μ_0 to a second probability measure μ_1. The optimality corresponds to minimizing a cost (the so-called Wasserstein distance) associated to the warping, which measures the effort needed to perform the corresponding motion. Informally, the effort is expressed as the cost it would require to move a pile of sand representing μ_0 toward a hole made of μ_1, by summing the cost (typically a squared distance) for each particle of sand to reach its destination in the hole (see Fig. 1). We refer to [32] for a review of the mathematical foundations of optimal transport.

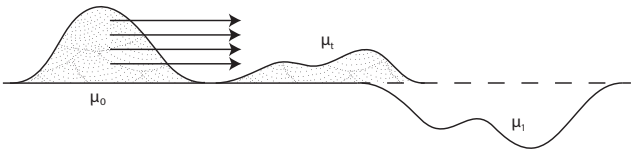


Fig. 1 The mass transportation problem consists in optimally moving a probability measure μ_0 represented by a pile of sand, toward a probability measure μ_1 making a hole. At an intermediate time $t \in [0,1]$, an interpolated probability measure μ_t, the *displacement interpolation*, is obtained. A *Wasserstein barycenter* generalizes this notion by considering more than 2 probability measures.

As a byproduct of the computation of this optimal transport, it is possible to define a geodesic μ_t, for $t \in [0,1]$ interpolating between the two input measures. This corresponds to the so-called displacement interpolation introduced by McCann [23]. Such an interpolation has several applications ranging from the analysis of PDEs to computer graphics, which we review below. Moreover, as introduced in [1], this interpolation between two densities can be extended to an arbitrary number of measures by defining a barycenter according to the transportation distance. However, a major bottleneck is the computational complexity of computing the optimal transport, geodesics and barycenters in arbitrary dimension. In this paper, we address these issues by leveraging the fact that these problems are easy to solve for 1-D distributions. We propose alternative definitions of barycenters using two frameworks based on 1-D projections of the measures. We describe the associated fast computational schemes, and show some applications in image processing (color transfer) and computer graphics (texture mixing).

N. Bonneel
Harvard University and LIRIS-CNRS
E-mail: nicolas.bonneel@liris.cnrs.fr

J. Rabin
GREYC, Université de Caen and CNRS

G. Peyré
CNRS and CEREMADE, Université Paris-Dauphine

H. Pfister
Harvard University

1.1 Previous work

Computational Optimal transport. There is a vast literature on the numerical computation and approximation of the optimal transport plan. For discrete measures (i.e. sums of Diracs), it boils down to the solution of a linear program, as initiated by Kantorovitch [21] which laid the modern foundations of transportation theory. There exist dedicated combinatorial optimization methods, such as the auction algorithm [5] and the Hungarian algorithm [18]. The L^2 optimal transport map is the solution of the celebrated Monge-Ampère nonlinear PDE. A variety of methods have been proposed to approximate numerically the solution to this equation, see for instance [4] and references therein.

Wasserstein geodesics. The Wasserstein geodesic (i.e. a minimizing length path interpolating between two measures) is easily computed by linearly interpolating between the identity and the optimal transport. It is thus a trivial by-product of the computation of the optimal map. Let us however notice that the landmark paper of Benamou and Brenier [3] proposes to actually proceed the other way around, i.e., to compute the geodesic as the solution of a convex optimization problem. The drawback of this approach is that it requires the addition of an extra dimension (time parameterizing the geodesic), but it allows the computation of an accurate approximation of the geodesic on a fixed discretization grid. This algorithm has recently been revisited using proximal splitting optimization schemes [25] ; we make use of this approach to compare the Wasserstein geodesics with the one obtained through our methods.

Wasserstein barycenters. Wasserstein barycenters generalize the notion of geodesic interpolation from two to an arbitrary number of measures. The mathematical foundation for the formulation of these barycenters (i.e. existence, uniqueness and linear programming formulation) is detailed in [1]. These barycenters have found application, for instance, in statistical estimation [6]. They enjoy an almost closed form expression in the case of Gaussian measures. This property is used in [15] to perform texture mixing of Gaussian texture models.

Cuturi and Doucet propose in [10] a numerical scheme to approximate the Wasserstein barycenter on an Eulerian grid. They smooth the Wasserstein distance using an entropic penalization, allowing them to perform a gradient descent. To reduce the numerical complexity of the barycenter computation, Rabin et al. [28] introduce a different variational problem that sums the Wasserstein distances of 1-D projections of the input measures. Our method generalizes the iterative 1-D histogram matching used in [26] to alter color palettes. Our work builds on the initial construction of Rabin et al. [28]. We propose a more formal exposition of this method and its main properties, and also present an alternative formulation based on the Radon transform.

Applications in imaging. There are numerous applications of mass transportation in image processing, computer vision and computer graphics. The Wasserstein distance leads to state-of-the-art results for several image retrieval problems, see for instance [30] for an early work on this topic. The optimal transport plan has been used for color transfer in images [26] and for meshing in computer graphics [14]. Displacement interpolation has been employed for image warping and registration [19,24], to remove flickering in old movies [12] and in computer graphics to perform manipulations on textures [22] and to interpolate reflectance for 3-D rendering [8]. The Wasserstein barycenter of Gaussian distributions has found applications for texture synthesis and mixing, using either non-parametric density estimations [28] and Gaussian density estimation [15].

1.2 Contributions

In this paper, we introduce two efficient methods to approximate the Wasserstein barycenter of an arbitrary number of measures based on 1-D projections. The first approach, that we call "Radon barycenter", computes 1-D barycenters of Radon projections of the input measures, and defines the resulting barycenter as a back-projection of these 1-D barycenters. This method leads to a fast numerical scheme for an Eulerian discretization of the measures (i.e. based on histograms on a regular lattice), using a discrete Radon transform. The second approach, that we call "sliced barycenter", is defined as the solution of an optimization problem which integrates the distances of all the Radon projections. A Lagrangian discretization (i.e. using point clouds with freely moving positions) is well adapted to the numerical resolution of a non-convex re-formulation of this optimization problem.

We demonstrate properties of these two barycenters, analyze their relationship and show how they compare in practice. We show that both approximations solve a similar variational problem that only differs in the lack of surjectivity of the Radon transform. We also prove that both barycenters exhibit similar translational and scaling properties as the exact Wasserstein barycenter at a fraction of its computational cost. We compare our approximation with the exact barycenter of two probability measures using a state of the art method [25]. We exemplify typical usages of these two complementary approaches to solve a problem of color harmonization in image processing, and a problem of texture mixing in computer graphics.

The code to reproduce the figure of this article is available online[1].

[1] `https://github.com/gpeyre/2014-JMIV-SlicedTransport`

1.3 Notations

We denote $\mathbb{S}^{d-1}$ the unit sphere in $\mathbb{R}^d$, and we define $\Omega^d = \mathbb{R}\times\mathbb{S}^{d-1}$. We denote $\mathrm{d}\theta$ the uniform measure on the sphere, which is normalized to satisfy $\int_{\mathbb{S}^{d-1}}\mathrm{d}\theta = 1$. We write $\mathscr{C}_0(X)$ the space of continuous functions on X tending to 0 at infinity, where in the following X is either $\mathbb{R}^d$ or Ω^d. It is a Banach space with respect to the norm

$$\forall f\in\mathscr{C}_0(X),\quad \|f\|_\infty = \max_{x\in X}|f(x)|.$$

We denote as $\mathscr{M}(X)$ the Radon measures on X, which is the space of finite Borel measures on X, and can also be represented as the dual of $\mathscr{C}_0(X)$, i.e., it is the space of continuous linear forms on $\mathscr{C}_0(X)$. We write

$$\forall(\mu,g)\in\mathscr{M}(X)\times\mathscr{C}_0(X),\quad \int_X g(x)\mathrm{d}\mu(x)\in\mathbb{R}$$

the duality pairing between these spaces, which evaluates at g the linear form defined by μ. $\mathscr{M}(X)$ is a Banach space with respect to the dual norm, which is the so-called total variation norm, $\forall\mu\in\mathscr{M}(X)$

$$\|\mu\|_{\mathrm{TV}} = \max\left\{\int_X g(x)\mathrm{d}\mu(x)\;;\; g\in\mathscr{C}_0(X),\ \|g\|_\infty\leqslant 1\right\}. \quad (1)$$

In the following, the convex cone of positive Radon measures is written

$$\mathscr{M}^+(X) = \left\{\mu\;;\;\forall f\in\mathscr{C}_0(\mathbb{R}^d), f\geqslant 0,\ \int f\mathrm{d}\mu\geqslant 0\right\}.$$

We denote as $\sharp$ the push-forward operator, which, for any measurable map $M: X\to Y$ defines a linear operator $M\sharp:\mathscr{M}(X)\to\mathscr{M}(Y)$ as, for any $\mu\in\mathscr{M}(X)$

$$\forall g\in\mathscr{C}_0(Y),\quad \int_Y g(y)\mathrm{d}(M\sharp\mu)(y) = \int_X g(M(x))\mathrm{d}\mu(x).$$

If $\mathrm{d}\mu(x)=\rho(x)\mathrm{d}x$ has a density ρ with respect to some measure $\mathrm{d}x$ (e.g., the Lebesgue measure on $\mathbb{R}^d$), and if M is a C^1 diffeomorphism, then one has

$$\mathrm{d}(M\sharp\mu)(y) = (\rho\circ M^{-1})(y)|\det(\partial M^{-1}(y))|\mathrm{d}y. \quad (2)$$

Using the disintegration theorem (see for instance [11]), one can slice a measure $\nu\in\mathscr{M}(\Omega^d)$ into its conditional measures with respect to the uniform measure on $\mathbb{S}^{d-1}$ to obtain a measure $\nu^\theta\in\mathscr{M}(\mathbb{R})$ for almost all $\theta\in\mathbb{S}^{d-1}$ outside a Borel set of zero measure, which satisfies, $\forall g\in\mathscr{C}_0(\Omega^d)$

$$\int_{\Omega^d} g(t,\theta)\mathrm{d}\nu(t,\theta) = \int_{\mathbb{S}^{d-1}}\left(\int_{\mathbb{R}} g(t,\theta)\mathrm{d}\nu^\theta(t)\right)\mathrm{d}\theta, \quad (3)$$

and such that for any Borel set $A\subset\mathbb{R}$, $\theta\in\mathbb{S}^{d-1}\mapsto\nu^\theta(A)\in\mathbb{R}$ is a Borel map.

The convex set of normalized positive probability measures is $\mathscr{M}_1^+(\mathbb{R}^d)\subset\mathscr{M}^+(\mathbb{R}^d)$, which are measures $\mu\in\mathscr{M}^+(\mathbb{R}^d)$ which satisfy $\mu(\mathbb{R}^d)=1$. We also denote $\bar{\mathscr{M}}_1^+(\Omega^d)$ the set of positive probability measures having normalized conditional mesures along the t variable, i.e.,

$$\bar{\mathscr{M}}_1^+(\Omega^d) = \left\{\nu\in\mathscr{M}_1^+(\Omega^d)\;;\;\forall\theta\in\mathbb{S}^{d-1},\quad \nu^\theta(\mathbb{R})=1\right\}$$

where $\nu^\theta\in\mathscr{M}_1^+(\mathbb{R})$ is the conditional measure defined according to the disintegration formula (3).

We denote as $\delta_x\in\mathscr{M}_1^+(\mathbb{R}^d)$ the Dirac measure at $x\in\mathbb{R}^d$, i.e.

$$\forall f\in\mathscr{C}_0(\mathbb{R}^d),\quad \int_{\mathbb{R}^d} f(y)\mathrm{d}(\delta_x)(y) = f(x).$$

We write $\mathscr{D}(X)$ the space of $\mathscr{C}^\infty(X)$ functions with compact support, and $\mathscr{D}^*(X)$ its dual, which is the space of distributions.

The Fourier transform of $f\in\mathrm{L}^1(\mathbb{R}^d)$ is defined as

$$\forall\omega\in\mathbb{R}^d,\quad \hat{f}(\omega) = \int_{\mathbb{R}^d} f(x)e^{-\mathrm{i}\langle\omega,x\rangle}\mathrm{d}x,$$

and the Fourier transform of a measure $\mu\in\mathscr{M}(\mathbb{R}^d)$ as

$$\forall\omega\in\mathbb{R}^d,\quad \hat{\mu}(\omega) = \int_{\mathbb{R}^d} e^{-\mathrm{i}\langle\omega,x\rangle}\mathrm{d}\mu(x).$$

Given a finite index set I, we define the simplex set of weights as

$$\Lambda_I = \left\{\lambda=(\lambda_i)_{i\in I}\in\mathbb{R}^I\;;\;\forall i\in I,\ \lambda_i\geqslant 0,\ \sum_{i\in I}\lambda_i = 1\right\} \quad (4)$$

where the notation $\mathbb{R}^I$ corresponds to the set of vectors indexed by I.

We define the following translation and scaling operators, for all $(s,u)\in\mathbb{R}^{+,*}\times\mathbb{R}^d$,

$$\forall x\in\mathbb{R}^d,\quad \varphi_{s,u}(x) = sx+u\in\mathbb{R}^d,$$
$$\forall(t,\theta)\in\Omega^d,\quad \psi_{s,u}(t,\theta) = (st+\langle u,\theta\rangle,\theta)\in\Omega^d.$$

We denote $\mathscr{O}(\mathbb{R}^d)$ the orthogonal group of $\mathbb{R}^d$, i.e. $\Phi:\mathbb{R}^d\mapsto\mathbb{R}^d$ is an invertible linear map with $\Phi^*=\Phi^{-1}$ the adjoint operator. For all $\Phi\in\mathscr{O}(\mathbb{R}^d)$ we denote

$$\tilde{\Phi}:(t,\theta)\in\Omega^d\mapsto(t,\Phi^*\theta)\in\Omega^d.$$

A measure μ is said to be radial (denoted $\mu\in\mathrm{Radial}(\mathbb{R}^d)$) if $\Phi\sharp\mu=\mu$ for all rotation $\Phi\in\mathscr{O}(\mathbb{R}^d)$. It is said to be centrally symmetric (denoted $\mu\in\mathrm{Central}(\mathbb{R}^d)$) if $S\sharp\mu=\mu$ for the central symmetry $S\in\mathscr{O}(\mathbb{R}^d)$ such that $S=-\mathrm{Id}_{\mathbb{R}^d}$.

2 Wasserstein Distance

2.1 Optimal Transport

For $(\mu_1,\mu_2)\in\mathscr{M}_1^+(\mathbb{R}^d)^2$, we define the L^2-Wasserstein distance $\mathrm{W}_{\mathbb{R}^d}(\mu_1,\mu_2)^2$ to be equal to

$$\inf\left\{\int_{\mathbb{R}^d\times\mathbb{R}^d}\|x_1-x_2\|^2\mathrm{d}\gamma(x_1,x_2)\ ;\ \gamma\in C(\mu_1,\mu_2)\right\} \tag{5}$$

where

$$C(\mu_1,\mu_2)=\left\{\gamma\in\mathscr{M}_1^+(\mathbb{R}^d\times\mathbb{R}^d)\ ;\ \Pi_i\sharp\gamma=\mu_i,\ i=1,2\right\}$$

where $\Pi_1(x_1,x_2)=x_1$ and $\Pi_2(x_1,x_2)=x_2$. We refer to [32] for more details regarding optimal transport and properties of the Wasserstein distance.

2.2 Wasserstein Barycenter on $\mathbb{R}^d$

Following [1], we define the Wasserstein barycenter as a natural extension of the variational formula for barycenters in $\mathbb{R}^d$.

Definition 1 (Wasserstein barycenter)**.** *Given $\lambda\in\Lambda_I$ and $(\mu_i)_{i\in I}\in\mathscr{M}_1^+(\mathbb{R}^d)^I$, we define*

$$\mathrm{Bar}^W_{\mathbb{R}^d}(\mu_i,\lambda_i)_{i\in I}=\operatorname*{argmin}_{\mu\in\mathscr{M}_1^+(\mathbb{R}^d)}\sum_{i\in I}\lambda_i\mathrm{W}_{\mathbb{R}^d}(\mu_i,\mu)^2. \tag{6}$$

Note that the variational problem is convex but it does not necessarily have a unique solution so that in general $\mathrm{Bar}^W_{\mathbb{R}^d}(\mu_i,\lambda_i)_{i\in I}$ is a (convex) set of measures. The solution can be shown to be unique (so that $\mathrm{Bar}^W_{\mathbb{R}^d}(\mu_i,\lambda_i)_{i\in I}$ is a singleton) if at least one of the μ_i does not give mass to so called "small sets" (sets of Hausdorff dimension strictly smaller than d), see [1]. A typical example of non-uniqueness can be shown on two input measures, for which a barycenter can be computed from any coupling measure γ solving (5), see [1], Section 4. If the two input measures are finite sums of Dirac's masses, then (5) is a finite dimensional linear program, which in general can fail to have a unique solution. The (convex) set of solution to this linear program thus defines a set of barycenters.

It is proved in [1] that this barycenter can be computed as the projection in $\mathbb{R}^d$ of a measure on $(\mathbb{R}^d)^I$ solving a linear program. This theorem shows that, in the particular case where the input measures are discrete probability measures (i.e. sums of weighted Diracs) then the barycenter measures solving (6) are discrete probability measures, which can be computed by solving a finite dimensional linear program. Note that since in this case all the input measures do give mass to small sets, then the barycenter can be non-unique for some degenerate configurations of input Diracs. Also note that solving such a high dimensional linear program is intractable for imaging applications. This is one of the main motivations to introduce alternative definitions of barycenters of measures.

The following proposition states some invariance properties of the Wasserstein barycenter with respect to translation, scaling, rotation and symmetry.

Proposition 1. *We consider $\lambda\in\Lambda_I$, $(\mu_i)_{i\in I}\in\mathscr{M}_1^+(\mathbb{R}^d)^I$. For all $(s,u)\in\mathbb{R}^{+,*}\times\mathbb{R}^d$,*

$$\mathrm{Bar}^W_{\mathbb{R}^d}(\varphi_{s,u}\sharp\mu_i,\lambda_i)_{i\in I}=\varphi_{s,u}\sharp\mathrm{Bar}^W_{\mathbb{R}^d}(\mu_i,\lambda_i)_{i\in I}, \tag{7}$$

and for all $\Phi\in\mathscr{O}(\mathbb{R}^d)$,

$$\mathrm{Bar}^W_{\mathbb{R}^d}(\Phi\sharp\mu_i,\lambda_i)_{i\in I}=\Phi\sharp\mathrm{Bar}^W_{\mathbb{R}^d}(\mu_i,\lambda_i)_{i\in I}. \tag{8}$$

In particular, one has

$$\forall i\in I,\mu_i\in\mathrm{Radial}(\mathbb{R}^d) \tag{9}$$

$$\Rightarrow\mathrm{Bar}^W_{\mathbb{R}^d}(\mu_i,\lambda_i)_{i\in I}\subset\mathrm{Radial}(\mathbb{R}^d), \tag{10}$$

and also

$$\forall i\in I,\mu_i\in\mathrm{Central}(\mathbb{R}^d) \tag{11}$$

$$\Rightarrow\mathrm{Bar}^W_{\mathbb{R}^d}(\mu_i,\lambda_i)_{i\in I}\subset\mathrm{Central}(\mathbb{R}^d). \tag{12}$$

The proof of this proposition, as well as all the other proofs of this section, can be found in Appendix A. The following proposition shows that the Wasserstein barycenter of translated and scaled copies of a given measure is also a translated and scaled copy.

Proposition 2. *We consider $\lambda\in\Lambda_I$, $\mu\in\mathscr{M}_1^+(\mathbb{R}^d)$. For all $(s_i,u_i)_{i\in I}\in(\mathbb{R}^{+,*}\times\mathbb{R}^d)^I$,*

$$\varphi_{s^\star,u^\star}\sharp\mu\in\mathrm{Bar}^W_{\mathbb{R}^d}(\varphi_{s_i,u_i}\sharp\mu,\lambda_i)_{i\in I},\quad\textit{where} \tag{13}$$

$$s^\star=\left(\sum_{i\in I}\lambda_i s_i^{-1}\right)^{-1}\quad\textit{and}\quad u^\star=\frac{\sum_{i\in I}\lambda_i s_i^{-1}u_i}{\sum_{i\in I}\lambda_i s_i^{-1}}. \tag{14}$$

2.3 Wasserstein Barycenter on $\mathbb{R}$

The following result shows that it is possible to compute a Wasserstein barycenter measure solving (6) in the 1-D case, with a close form expression. Note that if all the input measures contain Dirac atoms, the barycenter is not necessarily unique.

Proposition 3. *Let $\mu\in\mathscr{M}_1^+(\mathbb{R})$ be absolutely continuous with respect to the Lebesgue measure (i.e., such that μ has a density), and $(\mu_i)_{i\in I}\in\mathscr{M}_1^+(\mathbb{R})^I$. Denoting T_i such that $T_i\sharp\mu=\mu_i$ the optimal transport between μ and μ_i (which is unique), then*

$$\mu^\star=\left(\sum_{i\in I}\lambda_i T_i\right)\sharp\mu \tag{15}$$

is a barycenter measure solving (6)*, i.e. $\mu^\star\in\mathrm{Bar}^W_{\mathbb{R}}(\mu_i,\lambda_i)_{i\in I}$.*

For $\mu \in \mathscr{M}(\mathbb{R})$, we write the cumulative function as

$$\forall t \in \mathbb{R}, \quad C_\mu(t) = \mu(]-\infty, t]). \tag{16}$$

As for any non-decreasing function $f : \mathbb{R} \to \mathbb{R}$, one can define its pseudo inverse

$$\forall t \in \mathbb{R}, \quad f^+(t) = \inf\{s \in \mathbb{R} \; ; \; f(s) \geqslant t\}. \tag{17}$$

The following corollary shows that 1-D barycenters can be computed almost in closed form using inverse cumulative functions.

Corollary 1. *Given $(\mu_i)_{i\in I} \in \mathscr{M}_1^+(\mathbb{R})^I$, and $\lambda \in \Lambda_I$. Then*

$$\mu^\star = \frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_{i\in I} \lambda_i C^+_{\mu_i}(t)\right)^+, \tag{18}$$

where the derivative should be interpreted in the sense of distribution, satisfies $\mu^\star \in \mathrm{Bar}^W_{\mathbb{R}^d}(\mu_i, \lambda_i)_{i\in I}$, i.e., is a barycenter measure. In particular, it satisfies

$$C^+_{\mu^\star} = \sum_{i\in I} \lambda_i C^+_{\mu_i}.$$

2.4 Wasserstein Barycenter on Ω^d

We extend 1-D Wasserstein barycenters to barycenters of measures on Ω^d by essentially computing the barycenter along the t variable only. For this to be feasible, we restrict our attention to measures in $\bar{\mathscr{M}}_1^+(\Omega^d)$ having normalized conditional densities along the t variable.

Definition 2 (Wasserstein Barycenter on Ω^d)**.** *Given $(\nu_i)_{i\in I} \in \bar{\mathscr{M}}_1^+(\Omega^d)^I$ and $\lambda \in \Lambda_I$, we define the barycenters as*

$$\nu \in \mathrm{Bar}^W_{\Omega^d}(\nu_i, \lambda_i)_{i\in I} \in \bar{\mathscr{M}}_1^+(\Omega^d)$$

by, for almost all $\theta \in \mathbb{S}^{d-1}$,

$$\nu^\theta \in \mathrm{Bar}^W_{\mathbb{R}}(\nu_i^\theta, \lambda_i)_{i\in I}.$$

Considering the following extension of the Wasserstein distance to $\bar{\mathscr{M}}_1^+(\Omega^d)$ by integrating 1-D Wasserstein distances, $\forall (\nu_1, \nu_2) \in \bar{\mathscr{M}}_1^+(\Omega^d)^2$,

$$\mathrm{W}_{\Omega^d}(\nu_1, \nu_2)^2 = \int_{\mathbb{S}^{d-1}} \mathrm{W}_{\mathbb{R}}(\nu_1^\theta, \nu_2^\theta)^2 \mathrm{d}\theta,$$

we have the following characterization of the Wasserstein barycenter on Ω^d.

Proposition 4. *One has*

$$\mathrm{Bar}^W_{\Omega^d}(\nu_i, \lambda_i)_{i\in I} = \operatorname*{argmin}_{\nu \in \bar{\mathscr{M}}_1^+(\Omega^d)} \sum_{i\in I} \lambda_i \mathrm{W}_{\Omega^d}(\nu_i, \nu)^2. \tag{19}$$

The following proposition exposes some useful properties of barycenters in Ω^d.

Proposition 5. *If $\nu \in \bar{\mathscr{M}}_1^+(\Omega^d)$, then $\psi_{s,u}\sharp\nu \in \bar{\mathscr{M}}_1^+(\Omega^d)$, and*

$$\mathrm{Bar}^W_{\Omega^d}(\psi_{s,u}\sharp\nu_i, \lambda_i)_{i\in I} = \psi_{s,u}\sharp \mathrm{Bar}^W_{\Omega^d}(\nu_i, \lambda_i)_{i\in I} \tag{20}$$

$$\mathrm{Bar}^W_{\Omega^d}(\psi_{s_i,u_i}\sharp\nu, \lambda_i)_{i\in I} = \psi_{s^\star,u^\star}\sharp\nu \tag{21}$$

where $s^\star$ and $u^\star$ are defined in (14).

3 Radon Wasserstein Barycenters

Proposition 1 shows that it is computationally inexpensive to compute the Wasserstein barycenter of 1-D densities. It thus makes sense to seek for alternate definitions of barycenters of measures in $\mathbb{R}^d$ that rely on 1-D Wasserstein distances and barycenters. This section investigates a construction based on the Radon transform.

3.1 Radon Transform of Functions

We recall below classical definitions, and refer to [20] for more details. The Radon transform is first defined on integrable functions.

Definition 3 (Radon transform of functions)**.** *The Radon transform Rf of $f \in \mathrm{L}^1(\mathbb{R}^d)$ is defined as*

$$Rf(t, \theta) = \int_{\mathbb{R}^{d-1}} f(t\theta + U_\theta \gamma)\mathrm{d}\gamma \tag{22}$$

where $U_\theta \in \mathbb{R}^{d\times(d-1)}$ is any matrix such that its columns defines an orthogonal basis of $\theta^\perp$ (the hyperplane orthogonal to θ). This defines $R : \mathrm{L}^1(\mathbb{R}^d) \to \mathrm{L}^1(\Omega^d)$.

Its adjoint is defined on continuous functions as follows.

Definition 4 (Back-projection operator)**.** *The back projection R^*g of $g \in \mathscr{C}_0(\Omega^d)$ is defined as*

$$R^*g(x) = \int_{\mathbb{S}^{d-1}} g(\langle x, \theta\rangle, \theta)\mathrm{d}\theta.$$

This defines $R^* : \mathscr{C}_0(\Omega^d) \to \mathscr{C}_0(\mathbb{R}^d)$.

One has that R^*R is a translation invariant operator, i.e. a convolution

$$R^*Rf = h \star f \quad \text{where} \quad \hat{h}(\omega) = c\,\|\omega\|^{-(d-1)},$$

where $\star$ is the convolution on $\mathbb{R}^d$ and $c \in \mathbb{R}$ is a normalizing constant whose exact value depends on the dimension (see [20]). This relationship suggests a definition of a pseudo-inverse transform which operates on smooth functions so as to invert the low pass filter h.

Definition 5 (Inverse Radon transform of functions)**.** *The pseudo-inverse Radon transform R^+g of $g \in \mathscr{D}(\Omega^d)$ is defined as*

$$R^+g = h^+ \star (R^*g) \tag{23}$$

where h^+ is defined through $\hat{h}^+(\omega) = c^{-1}\|\omega\|^{d-1}$.

3.2 Radon Transform of Measures

Since R^* is defined on $\mathscr{C}_0(\mathbb{R}^d)$, the Radon transform is naturally extended to measures $\mu \in \mathscr{M}(\mathbb{R}^d)$ by duality as follows.

Definition 6 (Radon transform of measures)**.** *For all $\mu \in \mathscr{M}(\mathbb{R}^d)$, we set $\nu = R(\mu)$ be defined through, $\forall g \in \mathscr{C}_0(\Omega^d)$,*

$$\int_{\Omega^d} g(t,\theta)\mathrm{d}\nu(t,\theta) = \int_{\mathbb{R}^d} (R^* g)(x)\mathrm{d}\mu(x). \tag{24}$$

This defines $R : \mathscr{M}(\mathbb{R}^d) \to \mathscr{M}(\Omega^d)$.

The following proposition shows that the Radon transform of a measure gathers projections of the input measure along all possible directions.

Proposition 6. *For $\mu \in \mathscr{M}(\mathbb{R}^d)$, one has*

$$\forall\, \theta \in \mathbb{S}^{d-1}, \quad R(\mu)^\theta = P_\theta \sharp \mu$$

where $P_\theta : x \in \mathbb{R}^d \mapsto \langle x, \theta\rangle \in \mathbb{R}$,

and where $R(\mu)^\theta \in \mathscr{M}(\mathbb{R})$ is defined in (3)*.*

The proof of this proposition, as well as all the other proofs of this section, can be found in Appendix B.

The conditional measure ν^θ associated to $\nu \in \mathscr{M}(\Omega^d)$ is defined for almost all θ, i.e. on a Borel set of $\theta \in \mathbb{S}^{d-1}$ of measure 1. Proposition 6 shows that when $\nu = R(\mu)$, then ν^θ is in fact well defined for all $\theta \in \mathbb{S}^{d-1}$, because it is a push-forward measure.

We define in a way similar to Definition 6 the inverse Radon transform using the operator $R^{+,*} = R(R^*R)^{-1}$.

Definition 7 (Inverse Radon transform of measures)**.** *For all $\nu \in \mathscr{M}(\Omega^d)$, we set $\mu = R^+(\nu) \in \mathscr{D}^*(\mathbb{R}^d)$ be defined through, $\forall f \in \mathscr{D}(\mathbb{R}^d)$,*

$$\int_{\mathbb{R}^d} f(x)\mathrm{d}\mu(x) = \int_{\Omega^d} (R^{+,*} f)(t,\theta)\mathrm{d}\nu(t,\theta). \tag{25}$$

This defines $R^+ : \mathscr{M}(\Omega^d) \to \mathscr{D}^*(\mathbb{R}^d)$.

Note that for an arbitrary $\nu \in \mathscr{M}(\Omega^d)$ (i.e. not necessarily in the range $\mathrm{Im}(R)$ of R), $R^+\nu$ is a distribution and not necessarily a measure. One can however show that for $\nu = R(\mu) \in \mathrm{Im}(R)$, then $R^+(\nu) = \mu \in \mathscr{M}(\mathbb{R}^d)$ is a measure, as detailed in the following proposition. The proof of this proposition can be found in [7], Section 3.

Proposition 7. *$R : \mathscr{M}(\mathbb{R}^d) \to \mathscr{M}(\Omega^d)$ defined in* (24) *is injective, and $R^+R = \mathrm{Id}_{\mathscr{M}(\mathbb{R}^d)}$.*

The following lemma recapitulates useful commutation properties of the Radon transform with respect to translation and scaling.

Lemma 1. *One has, for $\mu \in \mathscr{M}_1^+(\mathbb{R}^d)$ and $\nu \in \mathscr{M}_1^+(\Omega^d)$, and for all $(s,u,\Phi) \in \mathbb{R}^{+,*} \times \mathbb{R}^d \times \mathscr{O}(\mathbb{R}^d)$,*

$$R(\varphi_{s,u}\sharp\mu) = \psi_{s,u}\sharp R(\mu) \tag{26}$$

$$R^+(\psi_{s,u}\sharp\nu) = \varphi_{s,u}\sharp R^+(\nu) \tag{27}$$

$$R(\Phi\sharp\mu) = \tilde{\Phi}\sharp R(\mu). \tag{28}$$

3.3 Radon Barycenter

According to Proposition 7, one has

$$R : \mathscr{M}_1^+(\mathbb{R}^d) \to R(\mathscr{M}_1^+(\mathbb{R}^d)) \subset \bar{\mathscr{M}}_1^+(\Omega^d),$$

although the inclusion on the right hand side is not an equality. This property allows us to define the Radon barycenter.

Definition 8 (Radon barycenter)**.** *Given $\lambda \in \Lambda_I$ and $(\mu_i)_{i\in I} \in \mathscr{M}_1^+(\mathbb{R}^d)^I$, we define*

$$\mathrm{Bar}^R_{\mathbb{R}^d}(\mu_i,\lambda_i)_{i\in I} = R^+\mathrm{Bar}^W_{\Omega^d}(R(\mu_i),\lambda_i)_{i\in I} \in \mathscr{D}^*(\mathbb{R}^d).$$

Since for $\nu \in \mathrm{Bar}^W_{\Omega^d}(R(\mu_i),\lambda_i)_{i\in I}$ one does not have in general $\nu \in \mathrm{Im}(R)$, $\mathrm{Bar}^R_{\mathbb{R}^d}(\mu_i,\lambda_i)_{i\in I}$ is composed of distributions and not necessarily measures.

The following proposition shows that the Radon barycenter enjoys the same invariance properties to scaling, translation and rotation as the classical Wasserstein barycenter.

Proposition 8. *Proposition 1 holds when replacing* $\mathrm{Bar}^W_{\mathbb{R}^d}$ *by* $\mathrm{Bar}^R_{\mathbb{R}^d}$.

The following proposition shows that, similarly to the usual Wassertstein barycenter, the Radon barycenter of translated and scaled copies of a given measure is also a translated and scaled copy.

Proposition 9. *Proposition 2 holds when replacing* $\mathrm{Bar}^W_{\mathbb{R}^d}$ *by* $\mathrm{Bar}^R_{\mathbb{R}^d}$.

3.4 Approximate Computation with Eulerian Discretization

Discretization grids. We consider here an Eulerian discretization of the Radon barycenter. This means that the considered measures in $\mathbb{R}^d$ are assumed to be discrete measures supported on the same grid of $N = n^d$ points in $\mathbb{R}^d$

$$\mathscr{G} = \{-n/2+1,\dots,n/2\}^d$$

(we assume for simplicity that n is even). Similarly, measures on Ω^d are also supported on a fixed grid

$$\tilde{\mathscr{G}} = \mathscr{T} \times \Theta = \{(t,\theta)\,;\, t \in \mathscr{T} \quad \text{and} \quad \theta \in \Theta\}$$

where $\mathscr{T} \subset \mathbb{R}$ and $\Theta \subset (-\pi,\pi]$ are finite sets.

Measures on grids. If X is a discrete set (which in the following will be either $\mathscr{G}$, $\tilde{\mathscr{G}}$ or $\mathscr{T}$), we denote

$$\forall\, a \in \mathbb{R}^X, \quad m_a^X = \sum_{x\in X} a_x \delta_x \in \mathscr{M}_1^+(X). \tag{29}$$

Following the notation introduced in (4), we denote Λ_X the set of normalized vectors

$$\Lambda_X = \left\{ a \in \mathbb{R}^X \,;\, \forall x \in X, a_x \geqslant 0 \quad \text{and} \quad \sum_{x\in X} a_x = 1 \right\}.$$

One thus has for $a \in \Lambda_X$, $m_a^X \in \mathscr{M}_1^+(X)$.

Discretized Wasserstein barycenter on $\mathscr{T}$. We first define approximate 1-D Wasserstein barycenters with an Eulerian discretization. The cumulative sum of $a \in \Lambda_{\mathscr{T}}$ is

$$\forall t \in \mathscr{T}, \quad I(a)_t = \sum_{t' \leqslant t} a_{t'}.$$

The cumulative distribution is defined by approximating with sums and interpolation the formula (16), for $\mu = m_a^{\mathscr{T}} \in \mathscr{M}_1^+(\mathbb{R})$

$$\forall t \in \mathbb{R}, \quad \bar{C}_\mu(t) = \mathrm{Interp}(I(a))(t).$$

Here, $\mathrm{Interp} : \mathbb{R}^{\mathscr{T}} \to \mathscr{C}_0(\mathbb{R})$ is an interpolation operator, that we take in the following to be piecewise linear. We then define the approximate barycenter on $\mathscr{T}$ of measures $(\mu_i = m_{a_i}^{\mathscr{T}})_{i \in I} \in \mathscr{M}_1^+(\mathbb{R})^I$ denoted

$$\mathrm{Bar}_{\mathscr{T}}(\mu_i, \lambda_i)_{i \in I} = m_{a^\star}^{\mathscr{T}}$$

by applying formula (18) on the grid $\mathscr{T}$, i.e.

$$\text{where} \quad \forall t \in \mathscr{T}, \quad a_t^\star = \frac{\mathrm{d}}{\mathrm{d}x}\left(\sum_{i \in I} \lambda_i \bar{C}_{\mu_i}^+(x) \right)^+ (t).$$

In practice, this formula is computed accurately by computing the inverse cumulative function on a uniform grid of $[0,1]$ of the same granularity as the spatial discretization, and computing the derivative with finite differences on this grid.

Discretized Wasserstein barycenter on $\tilde{\mathscr{G}}$. One computes Eulerian barycenters on Ω^d by computing 1-D barycenters of the marginals restricted to the grid $\tilde{\mathscr{G}}$. Indeed, we have for $\beta \in \Lambda_{\tilde{\mathscr{G}}}$, denoting $\nu = m_\beta^{\tilde{\mathscr{G}}}$, the disintegration formula on the grid

$$\forall \theta \in \Theta, \nu^\theta = m_{\beta_{\cdot,\theta}}^{\mathscr{T}} \quad \text{where} \quad \beta_{\cdot,\theta} = (\beta_{(t,\theta)})_{t \in \mathscr{T}} \in \mathbb{R}^{\mathscr{T}}.$$

The approximate barycenter on $\tilde{\mathscr{G}}$ of measures $(\nu_i = m_{\beta_i}^{\tilde{\mathscr{G}}})_{i \in I} \in \mathscr{M}_1^+(\Omega^d)^I$ is thus

$$\mathrm{Bar}_{\tilde{\mathscr{G}}}(\nu_i, \lambda_i)_{i \in I} = m_{\beta^\star}^{\tilde{\mathscr{G}}} = \nu^\star$$
$$\text{where} \quad \forall \theta \in \Theta, \quad (\nu^\star)^\theta = \mathrm{Bar}_{\mathscr{T}}(\nu_i^\theta, \lambda_i)_{i \in I}.$$

Discrete Radon transform In the following, we investigate the use of the Fast Slant Stack Radon transform [2]. It has the property to faithfully approximate the geometry of the Radon transform, i.e., it exactly computes integrals over 1-D rays for band limited functions. Note that other discretizations could be used as well, see for instance [9]. In the case of a 2-D Fast Slant Stack transform, the sampling grid $\tilde{\mathscr{G}}$ is recto-polar (so that $\tilde{\mathscr{G}}$ is in fact not an exactly equi-spaced grid, but we ignore this technicality here) and $|\mathscr{T}| = n$, $|\Theta| = 4n$. This Fast Slant Stack implements both the computation of the Radon transform and its adjoint with fast algorithms. These algorithms assume that the data is sampled from a band limited function, faithfully integrated using Shannon interpolation. This can thus result in negative values in the Radon transform, and in turn necessitates a careful implementation of the barycenter computation.

We thus assume that we have at our disposal a discrete Radon transform (in our case the Fast Slant Stack), which is a linear map $\tilde{R} : \mathbb{R}^{\mathscr{G}} \mapsto \mathbb{R}^{\tilde{\mathscr{G}}}$, and also have access to its adjoint $\tilde{R}^* : \mathbb{R}^{\tilde{\mathscr{G}}} \mapsto \mathbb{R}^{\mathscr{G}}$. The Moore-Penrose pseudo-inverse

$$\tilde{R}^+(\beta) = (\tilde{R}^*\tilde{R})^{-1}\tilde{R}^*(\beta) = \operatorname*{argmin}_\alpha \|\tilde{R}\alpha - \beta\|$$

is usually computed by a conjugate gradient descent. As reported in [2], it is possible to introduce a simple pre-conditionner for the Fast Slant Stack inversion that accelerates convergence of the conjugate descent, and is a major computational advantage for this approach.

This discrete Radon transform allows one to approximate the Radon transform of measures defined in (24) as

$$\forall \alpha \in \mathbb{R}^{\mathscr{G}}, \quad R(m_\alpha^{\mathscr{G}}) \approx m_{\tilde{R}(\alpha)}^{\tilde{\mathscr{G}}}.$$

We leave for future work the theoretical analysis of this approximation when $m_\alpha^{\mathscr{G}} \to \mu$ and (N,P) increases toward $+\infty$.

Approximated Radon Barycenters Making use of these discrete constructions (barycenters on $\tilde{\mathscr{G}}$ and Radon transform on $\mathscr{G}$), we are now ready to define the approximate Eulerian barycenter of measures supported on $\mathscr{G}$. We are thus given as input Eulerian discretized densities

$$\forall i \in I, \quad \mu_i = m_{\alpha_i}^{\mathscr{G}} \quad \text{where} \quad \alpha_i \in \mathbb{R}^{\mathscr{G}}.$$

The algorithm then computes the discretized Radon transform

$$\forall i \in I, \quad \beta_i = \tilde{R}(\alpha_i) \in \mathbb{R}^{\tilde{\mathscr{G}}}.$$

For any $\lambda \in \Lambda_I$, our Eulerian discretized Radon barycenter

$$\mathrm{Bar}_{\mathscr{G}}^R(\mu_i, \lambda_i)_{i \in I} = m_{\alpha^\star}^{\mathscr{G}} \quad \text{where} \quad \begin{cases} \alpha^\star = \tilde{R}^+\beta^\star, \\ m_{\beta^\star}^{\tilde{\mathscr{G}}} = \mathrm{Bar}_{\tilde{\mathscr{G}}}(m_{\beta_i}^{\tilde{\mathscr{G}}}, \lambda_i)_{i \in I}. \end{cases}$$

This barycenter is hence intended to approximate an element of $\mathrm{Bar}_{\mathbb{R}^d}^R(\mu_i)_{i \in I}$, with the constraint of being supported on $\mathscr{G}$.

4 Sliced Wasserstein Barycenter

4.1 Sliced Wasserstein Barycenter

Following [28] which defines a sliced barycenter of discrete measures, we consider here a similar sliced variational formulation for arbitrary measures. We first define the sliced Wasserstein distance as

$$\mathrm{SW}_{\mathbb{R}^d}(\mu_1, \mu_2)^2 = \mathrm{W}_{\Omega^d}(R\mu_1, R\mu_2)^2 \tag{30}$$
$$= \int_{\mathbb{S}^{d-1}} \mathrm{W}_{\mathbb{R}}(P_\theta \sharp \mu_1, P_\theta \sharp \mu_2)^2 \mathrm{d}\theta. \tag{31}$$

where we remind that $\mathrm{d}\theta$ is the uniform measure on $\mathbb{S}^{d-1}$, normalized so that $\int_{\mathbb{S}^{d-1}} \mathrm{d}\theta = 1$.

Definition 9 (Sliced Wasserstein Barycenter)**.** *Given $\lambda \in \Lambda_I$ and $(\mu_i)_{i\in I} \in \mathscr{M}_1^+(\mathbb{R}^d)^I$ we define*

$$\mathrm{Bar}^S_{\mathbb{R}^d}(\mu_i,\lambda_i)_{i\in I} = \operatorname*{argmin}_{\mu\in\mathscr{M}_1^+(\mathbb{R}^d)} \sum_i \lambda_i \mathrm{SW}_{\mathbb{R}^d}(\mu_i,\mu)^2. \tag{32}$$

4.2 Comparison of Radon and Sliced Barycenters

The following proposition compares the variational formulations of the Radon and sliced Wasserstein barycenters.

Proposition 10. *Denoting*

$$\mathscr{E}(v) = \sum_{i\in I} \lambda_i \mathrm{W}_{\Omega^d}(R\mu_i, v)^2, \tag{33}$$

one has

$$\mathrm{Bar}^R_{\mathbb{R}^d}(\mu_i,\lambda_i)_{i\in I} = R^+ \operatorname*{argmin}_{\bar{\mathscr{M}}_1^+(\Omega^d)} \mathscr{E}, \tag{34}$$

$$\mathrm{Bar}^S_{\mathbb{R}^d}(\mu_i,\lambda_i)_{i\in I} = R^+ \operatorname*{argmin}_{\bar{\mathscr{M}}_1^+(\Omega^d)\cap \mathrm{Im}(R)} \mathscr{E}. \tag{35}$$

The proof of this proposition, as well as all the other proofs of this section, can be found in Appendix C. The following proposition shows that the sliced barycenter enjoys the same invariance properties as the Radon barycenter.

Proposition 11. *Proposition 1 holds when replacing* $\mathrm{Bar}^W_{\mathbb{R}^d}$ *by* $\mathrm{Bar}^S_{\mathbb{R}^d}$.

Proposition 12. *Proposition 2 holds when replacing* $\mathrm{Bar}^W_{\mathbb{R}^d}$ *by* $\mathrm{Bar}^S_{\mathbb{R}^d}$.

4.3 Sliced Barycenter with Lagrangian Discretization

Directly solving the variational problem (32) is intractable for any realistic application. Indeed, even for discrete input measures, the barycenter might not be in general discrete. Instead, we consider a numerical scheme that performs the optimization of (32) over the (non-convex) set of discrete sums of Diracs. We parameterize a discrete measure with equal weights as

$$\mu_X = \frac{1}{N}\sum_{k=1}^N \delta_{X_k} \tag{36}$$

where $X = (X_k)_{k=1}^N \in \mathbb{R}^{d\times N}$ and $X_k \in \mathbb{R}^d$.

Given a set $(\mu_i)_{i\in I}$ of discrete input measures, i.e. $\mu_i = \mu_{X^{(i)}}$ for $X^{(i)} \in \mathbb{R}^{d\times N}$, we consider the following non-linear program to approximate solutions of (32)

$$\min_{X\in\mathbb{R}^{d\times N}} \left\{ \mathscr{E}(X) = \sum_{i\in I} \frac{\lambda_i}{2} \mathrm{SW}_{\mathbb{R}^d}(\mu_{X^{(i)}}, \mu_X)^2 \right\}. \tag{37}$$

The following theorem shows that this energy is smooth, which contrasts with the same energy defined with the usual Wasserstein distance $\mathrm{W}_{\mathbb{R}^d}$ instead of $\mathrm{SW}_{\mathbb{R}^d}$.

Theorem 1. *$\mathscr{E} : \mathbb{R}^{d\times N} \to \mathbb{R}$ is a $\mathscr{C}^1$ function with a uniformly Lipschtiz gradient. Its gradient at $X \in \mathbb{R}^{d\times N}$ with distinct points reads*

$$\nabla\mathscr{E}(X) = \sum_{i\in I} \lambda_i \int_{\mathbb{S}^{d-1}} (X_\theta - X^{(i)}_\theta \circ \sigma_{X^{(i)}_\theta} \circ \sigma_{X_\theta})\theta \,\mathrm{d}\theta \tag{38}$$

where $X_\theta = (\langle X_i, \theta\rangle)_{i=1}^N \in \mathbb{R}^N$ and for any $Y \in \mathbb{R}^N$, σ_Y is any permutation (which is not necessarily unique) of $\{1,\dots,N\}$ which orders the values in Y, i.e.

$$Y_{\sigma(1)} \leqslant Y_{\sigma(2)} \leqslant \dots \leqslant Y_{\sigma(N)}.$$

The proof of this theorem can be found in Appendix D. Problem (37) is non-convex, and one computes a stationary point (in practice a local minimum) through a gradient descent

$$X^{[\ell+1]} = X^{[\ell]} - \tau_\ell \nabla\mathscr{E}(X^{[\ell]}) \tag{39}$$

with a given initialization $X^{[0]}$, and where $\nabla\mathscr{E}(X^{[\ell]})$ is computed using (38), and τ_ℓ is a gradient step size. Choosing $0 < \tau < \tau_\ell < 2/\kappa$ ensures convergence, where $\kappa > 0$ is the uniform Lipschitz constant of $\nabla\mathscr{E}$. Note that the constant κ depends on the input point clouds $(X^{(i)})_{i\in I}$, and we found in practice that κ is close to 1, see also the proof in Appendix D for more insights about this.

In order to implement numerically the iterations (39), one discretizes the set of directions. It corresponds to the use of a finite set $\Theta \subset \mathbb{S}^{d-1}$, and a minimization of the energy

$$\mathscr{E}_\Theta(X) = \sum_{i\in I} \frac{\lambda_i}{2|\Theta|} \sum_{\theta\in\Theta} \mathrm{W}_{\mathbb{R}}(P_\theta\sharp\mu_{X^{(i)}}, P_\theta\sharp\mu_X)^2.$$

While this function is not $\mathscr{C}^1$ on the whole space $\mathbb{R}^{d\times N}$, it is differentiable (and in fact quadratic) almost everywhere. At a point where it is differentiable, one can use formula (38), where the integral $\int_{\mathbb{S}^{d-1}}$ is replaced by a finite sum $\sum_\Theta$. The gradient descent (39) is advantageously replaced by a Newton descent

$$X^{[\ell+1]} = X^{[\ell]} - H_\ell^{-1}\nabla\mathscr{E}(X^{[\ell]}) \tag{40}$$

where

$$H_\ell = \nabla^2\mathscr{E}(X^{[\ell]}) = \frac{1}{|\Theta|}\sum_{\theta\in\Theta} \theta\theta^* \in \mathbb{R}^{d\times d}$$

is the Hessian matrix of $\mathscr{E}$ (which thus does not depends on ℓ). In 2-D, we use a set of $|\Theta|$ directions equi-spaced on the circle, in which case $H_\ell = \frac{1}{2}\mathrm{Id}_{2\times 2}$. In higher dimensions $d > 2$, we use random directions drawn uniformly on $\mathbb{S}^{d-1}$, and one can show that H_ℓ converges almost surely to $\frac{1}{d}\mathrm{Id}_{d\times d}$, so that in practice one can use this matrix in place of H_ℓ

in (40). Although we observed that this approximated Newton scheme works well in our numerical simulation, it is not possible to give theoretical claim about its convergence speed, since the underlying function is not twice differentiable, and the Hessian matrix is computed with some error when using $\frac{1}{d}\mathrm{Id}_{d\times d}$.

4.4 Sliced Transport with Lagrangian Discretization

Beside the computation of barycenters, the sliced Wasserstein distance (30) can be used to approximate the transportation map from a given density $\mu_{X^{[0]}}$ toward a second density μ_Y, for $(X^{[0]},Y)\in(\mathbb{R}^{d\times N})^2$. This application was initially introduced by Marc Bernot and first presented in [28] for applications to texture synthesis.

We obtain this map by following the descent flow of the energy

$$\forall X\in\mathbb{R}^{d\times N},\quad \mathscr{F}_Y(X)=\frac{1}{2}\mathrm{SW}_{\mathbb{R}^d}(\mu_X,\mu_Y)^2$$

initialized from $X^{[0]}$, which can be formally written as the flow $t\mapsto X_t\in\mathbb{R}^{d\times N}$ defined by the PDE

$$\forall t>0,\quad \frac{\partial X_t}{\partial t}=-\nabla\mathscr{F}_Y(X_t) \tag{41}$$

with $X_0=X^{[0]}$ at time $t=0$. Note that the gradient of $\mathscr{F}_Y$ is given by Theorem 1 in the case of a single input density, i.e., $|I|=1$.

In order to numerically approximate the flow (41), we discretize the time dimension using an explicit Euler scheme (which corresponds to a gradient descent) and the set of directions used in the definition of $\mathrm{SW}_{\mathbb{R}^d}$. In order for the flow to converge to a stationary point of $\mathscr{F}_Y$, we use a stochastic gradient descent. At each iteration ℓ, we consider a finite number of orientations $\Theta_\ell\subset\mathbb{S}^{d-1}$ drawn uniformly at random. Defining the partial energy

$$\mathscr{F}_Y^\ell(X)=\frac{1}{|\Theta_\ell|}\sum_{\theta\in\Theta_k}\mathrm{W}_{\mathbb{R}}(P_\theta\sharp\mu_{X^{(i)}},P_\theta\sharp\mu_X)^2,$$

one step of the stochastic gradient descent is defined as

$$X^{[\ell+1]}=X^{[\ell]}-\tau_\ell\nabla\mathscr{F}_Y^\ell(X^{[\ell]}) \tag{42}$$

where $\tau_\ell>0$ is a step size.

We denote

$$X^\star\in\lim_{\ell\to+\infty}X^{[\ell]} \tag{43}$$

any limiting point cloud in the adherence of the sequence of iterates. Since this sequence is bounded by coercivity of $\mathscr{F}_X$, such a point cloud always exists.

Note that the color transfer method introduced by Pitié et al. [26] corresponds to the iterations (42) when using $|\Theta_\ell|=3$ randomized orthogonal directions at each step. Figure 3 in the next section shows that using more directions improves the visual quality of the result.

Experimentally, as detailed in Section 5.3, we make the following crucial observations.

Remark 1. The step size τ_ℓ can be set constant, i.e. $\forall\ell,\tau_\ell=\tau$, and the iterates always converge toward a local minimum of $\mathscr{F}_Y$. A heuristic explanation for this observed property is that, at a global minimum X of $\mathscr{E}_Y(X)$, for all $\theta\in\mathbb{S}^{d-1}$, each term $\mathrm{W}_{\mathbb{R}}(P_\theta\sharp\mu_{X^{(i)}},P_\theta\sharp\mu_X)^2$ also reaches its global minimum. For a convex energy, this property is known to imply convergence of stochastic gradient descent with a fixed step size, see [31].

Remark 2. All local minima of $\mathscr{F}_Y$ appear to be global minima. Although we have no formal proof of this phenomenon, it is illustrated on measures made of two Diracs in Section 5.1. This implies that $X^\star$ is a global minimum of $\mathscr{F}_Y$, hence $\mathscr{F}_Y(X^\star)=0$ and

$$\mu_{X^\star}=\mu_Y \tag{44}$$

i.e. the measure $\mu_{X^{[\ell]}}$ converges (in the weak-* topology of Radon measures) toward μ_Y.

The sliced transport map $T^S:\mathbb{R}^d\mapsto\mathbb{R}^d$ is defined on the support of $\mu_{X^{[0]}}$ as

$$\forall k\in\{1,\dots,N\},\quad T^S(X_k^{[0]})=X_k^\star. \tag{45}$$

The (empirically observed) property (44) ensures that T^S satisfies $T^S\sharp\mu_{X^{[0]}}=\mu_Y$, i.e., T^S is a valid transport plan between the measures.

5 Numerical Illustrations

We emphasize that this paper introduces two different approaches (Sliced and Radon) together with their corresponding discretization (Lagrangian and Eulerian) to cope with the variety of image processing and computer graphics applications that optimal transport is targeting. This section compares these two methods on synthetic examples and explores a few of these applications in order to illustrate the relative benefit of each method.

5.1 A Case Study: Sliced Wasserstein Distance for Pairs of Diracs

As discussed in Section 4.4, we empirically found that the sliced Wasserstein distance $X\mapsto\mathrm{SW}_{\mathbb{R}^d}(\mu_X,\mu_Y)$ to a given measure μ_Y for $Y\in\mathbb{R}^{N\times d}$ has no local minimum, i.e., only has global minima satisfying $\mu_X=\mu_Y$, saddle points, and local maxima. This is of primary importance because in practice a descent scheme avoids saddle points and local maxima (since these are unstable stationary point of the flow),

and the gradient flow (41) converges to a global minimum, which in turn defines an assignment.

While we do not give a formal proof of this statement, we illustrate this point on a simple example in 2-D (i.e. $d = 2$) with point clouds having two masses (i.e. $N = 2$). We fix

$$\begin{cases} Y = \{Y_1, Y_2\} = \{(0,-1),(0,+1)\} \quad \text{and} \\ X(u) = \{X_1, X_2\} = \{u, -u\}, \end{cases}$$

and only let $u = (x,y) \in \mathbb{R}^2$ varies (see Figure 2 for an illustration). We then compare the Wasserstein and Sliced Wasserstein distances

$$\forall u \in \mathbb{R}^2, \quad \begin{cases} \mathscr{E}^W(u) = \mathrm{W}_{\mathbb{R}^2}(\mu_{X(u)}, \mu_Y)^2, \\ \mathscr{E}^S(u) = \mathrm{SW}_{\mathbb{R}^2}(\mu_{X(u)}, \mu_Y)^2, \\ \mathscr{E}^S_\Theta(u) = \frac{1}{|\Theta|} \sum_{\theta \in \Theta} \mathrm{W}_{\mathbb{R}}(P_\theta \sharp \mu_{X(u)}, P_\theta \sharp \mu_Y)^2. \end{cases}$$

After some calculations, we get the following expressions for $\mathscr{E}^W$ and $\mathscr{E}^S$

$$\forall (x,y) \in \mathbb{R}^2, \begin{cases} \mathscr{E}^W(x,y) = 2\left(x^2 + (|y|-1)^2\right), \\ \mathscr{E}^S(x,y) = x^2 + y^2 + 1 - \frac{4}{\pi}\left(x + y \cdot \operatorname{atan}\left(\frac{y}{x}\right)\right) \end{cases}$$

while $\mathscr{E}^S_\Theta(u)$ is evaluated numerically using a discrete set of orientations Θ. Figure 2 shows a comparison of these two distances. We can see that the Sliced Wasserstein distance (as well as the Wasserstein distance) has no local minimum, although there are three saddle points at $u = (0,0)$ and $u = \pm(\frac{2}{\pi}, 0)$, which separate two basins of attraction associated to the two global minima.

5.2 Numerical Considerations for the Sliced Transport

Influence of the number of directions. We first illustrate the special case discussed in Section 4.4 of the transport of a discrete distribution (a sum of Dirac masses) μ_0 toward another, μ_1. This boils down to an assignment problem. We resort to the stochastic gradient descent detailed in (42) to compute a Sliced Wasserstein transport T^S. This map T^S always numerically verifies $T^S \# \mu_0 = \mu_1$. Nevertheless, it can be far from the optimal Wasserstein transport map T^W when using a small number of directions at each iteration. We illustrate this in Figure 3, that shows the distributions obtained when interpolating the transport map T^S, that is, we compute $\mu^S_\lambda = [(1-\lambda)\mathrm{Id} + \lambda T^S] \sharp \mu_0$ for $\lambda \in [0,1]$, when varying the number of directions Θ_ℓ used at each iteration. Using more sampling directions tends to provide more regular transport maps. However, we note that the Sliced Wasserstein transport can provide a different assignment from the optimal Wasserstein map, even when using a large set of directions (Fig. 3, third example).

Influence of local minima. Since the algorithm detailed in Section 4.3 performs a non-convex energy minimization (see (37)), it is important to understand the influence of the initialization of the descent. Figure 4 analyzes on a simple example the effect of the presence of local minima. The center plots (b) and (c) each show two results (blue and red dots) approximating the sliced iso-barycenter $\mu^S_{1/2}$ using our non-convex gradient descent, as well as the Wasserstein barycenter $\mu^W_{1/2}$ (black dots) computed via linear programming since there are only two distributions. Each result is obtained using a random initialization with samples independently drawn from an isotropic Gaussian having the same mean and variance as μ_0. While this clearly shows that different initializations lead to different estimates, this also shows that the impact of the initialization is quite modest.

5.3 Numerical Comparison of the Barycenters

This section compares the following barycenters in a 2-D ($d = 2$) setting:

- The original Wasserstein barycenter $\mathrm{Bar}^W_{\mathbb{R}^d}$ (see Definition 1), which can only be computed numerically for 2 distributions, i.e. $|I| = 2$, and thus corresponds to the Wasserstein geodesic between the two measures. We use the proximal splitting method of [25] to estimate this barycenter with a Eulerian discretization on a fixed grid.
- The Radon barycenter $\mathrm{Bar}^R_{\mathbb{R}^d}$ (see Definition 8). It is approximated with the numerical scheme presented in Section 3.4 with an Eulerian discretization.
- The sliced barycenter $\mathrm{Bar}^S_{\mathbb{R}^d}$ (see Definition 9). It is approximated with the numerical scheme presented in Section 4.3 with a Lagrangian discretization. If not stated otherwise, we use $|\Theta| = 10$ directions uniformly sampled on the half circle.

Comparison of the Sliced, Radon and Wasserstein Geodesics. In general, the sliced and Radon barycenters differ from the original Wasserstein barycenter. While the Wasserstein barycenter of Gaussian distributions is always a Gaussian distribution [1], Figure 5 shows that this is not the case for the Radon barycenter when the Gaussians are not isotropic.

Figure 6 shows a more detailed comparison of both smooth (Gaussian mixture) and non-smooth (characteristic function of animal-like shapes) densities. Only the edge of the barycentric triangle is available for the Wasserstein barycenter, since there is no efficient algorithm to approximate the Wasserstein barycenter of more than two measures.

Comparison of the Sliced and Radon barycenters. As emphasized by Proposition 10, while $\mathrm{Bar}^R_{\mathbb{R}^d}$ and $\mathrm{Bar}^S_{\mathbb{R}^d}$ are mathematically different, this difference is rather small, and is

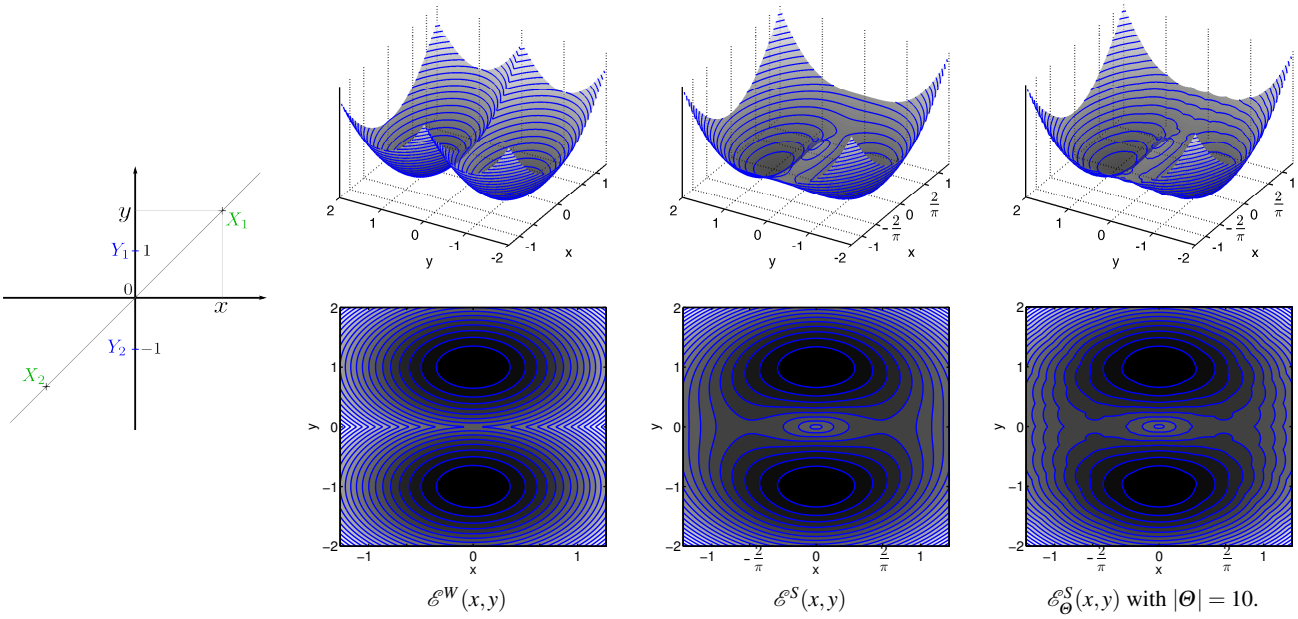


Fig. 2 Comparison of $\mathrm{SW}^2_{\mathbb{R}^2}$ and $\mathrm{SW}^2_{\mathbb{R}^2}$, and its numerical approximation using 10 directions, as an elevation surface (top row) and its corresponding 2d map (bottom row).

solely due to the lack of surjectivity of the Radon transform. We numerically evaluated this difference by computing

$$\|R(\mathrm{Bar}^R_{\mathbb{R}^d}(\mu_i,\lambda_i)_{i\in I}) - \mathrm{Bar}^W_{\Omega^d}(R(\mu_i),\lambda_i)_{i\in I}\|_{\mathrm{TV}},$$

where $\|\cdot\|_{\mathrm{TV}}$ is the total variation of the measure defined in (1) and corresponds to the L^1 norm of the density in the case of an absolutely continuous measure. This measures the relative error due to the lack of surjectivity of R. Among several sets of discretized measures μ_i and weights λ_i, this relative error remained at approximately 0.15%. This said, the main difference between the sliced and Radon barycenter lies in their discretizations: $\mathrm{Bar}^R_{\mathbb{R}^d}$ is approximated with an Eulerian scheme and $\mathrm{Bar}^S_{\mathbb{R}^d}$ with a Lagrangian scheme.

Figure 6 shows that the discretized barycenters are quite similar when computing the barycenter of three measures. Figure 8 shows a similar comparison for the iso-barycenter of four measures. Figure 7 shows what could be considered as a failure of the method to adapt to the computation of complex image barycenters.

Comparison of computational complexity. A typical Radon barycenter of three two-dimensional pdfs discretized on a 1024×1024 pixel grid, and the principled Fast Slant Stack Radon transform with 2048 slices, requires 11 seconds to precompute the initial Radon transforms, and 170 seconds to compute 32 Radon barycenters, with unoptimized parallel Matlab code. It is possible to accelerate this timing using less precise Radon transform. For instance, using Matlab's implementation of the Radon transform with 180 slices requires 14 seconds to compute these 32 barycenters on a single core. In comparison with the Eulerian proximal splitting method of Papadakis et al. [25], the Wasserstein barycenter between two 1024×1024 distributions with 32 time steps and $100,000$ iterations to achieve an acceptable convergence requires on average 72 hours, using an optimized C++ vectorized and parallel implementation (see Fig. 6 for a display of the resulting barycenters).

A sliced barycenter of three distributions, each approximated with 40k Dirac masses and 100 directions, requires 140 seconds using 100 iterations of Newton descent or 18 seconds using the stochastic Newton descent with subsets of 10 directions. With a finer set of 1000 directions and the same setup, the stochastic Newton descent with subsets of 100 directions requires 168 seconds.

Comparison with the entropy regularized barycenter [10]. Cuturi and Doucet proposed in [10] a method to approximate the Wasserstein barycenter on a fixed grid, hence using the Eulerian discretization presented in Section 3.4. Their method performs a gradient descent on a smoothed Wasserstein distance. This smoothing is obtained by adding an entropic penalization to the linear cost function (5) defining the transportation distance. Figure 9 shows a visual comparison of the iso-barycenters computed with this approach as well as with the sliced and Radon methods.

The result obtained with the method of [10] is produced in two hours on a GPU, using a 150×150 sampling grid. In contrast, our Radon barycenter computed on a grid of 400×400 pixels (which is zero-padded to 1200×1200 pixels to avoid Radon transform artifacts) is obtained in 40 seconds using the fast slant stack approach with 1200 directions, and

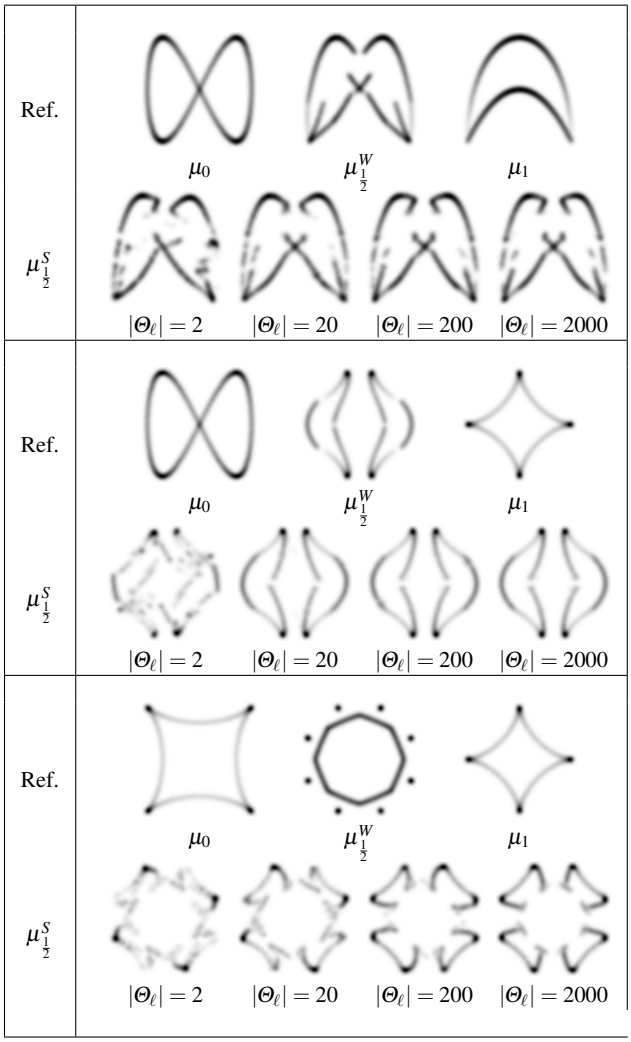


Fig. 3 We consider the optimal Wasserstein barycenter $\mu_{\frac{1}{2}}^W$ between μ_0 and μ_1. We show the Sliced Wasserstein interpolation $\mu_{\frac{1}{2}}^S$ using our stochastic Newton descent (42) with different number of directions $|\Theta_\ell|$. The density is displayed using a Parzen density estimation.

2 seconds with Matlab built-in Radon transforms with 180 directions, on a single core of a laptop. Similarly, our Sliced barycenter implemented in Matlab produced the interpolation in 20 minutes, using 4×10^5 points sampled on a interpolated grid of 1000×1000 pixels, with 100 directions and 100 iterations.

5.4 Application to Texture Mixing

To illustrate the usefulness of the Radon barycenter, we apply it to the problem of texture mixing. The Radon barycenter is well suited to this application which requires an Eulerian discretization in order to interpolate power-spectra computed on the uniform grid of Fourier frequencies. This would be hardly feasible using the Lagrangian discretization of the Sliced barycenter.

Texture mixing. Given a set of input texture images $\{f^{[i]}\}_{i\in I}$, where each $f^{[i]} \in \mathbb{R}^N$ is a grayscale image of $N = n \times n$ pixels, the goal of texture mixing is to produce a set of random vectors $\{F^{[i]}\}_{i\in I}$, and an interpolation method $\lambda \in \Lambda_I \mapsto F_\lambda$. In particular, it means that if λ is 0 excepted at the i^{th} coordinate, then $F_\lambda = F^{[i]}$ (interpolation at the vertices of the simplex indexed by I). Texture mixing is a generalization of texture synthesis (which simply corresponds to the case $|I| = 1$), in the sense that any realization $\tilde{f}^{[i]}$ of the random vector $F^{[i]}$ should look both "random" and visually similar (but not equal) to the input $f^{[i]}$.

Spot-noise (SN) texture model. Following the work of [16] (which introduces the name "spot noise" model), we consider stationary Gaussian random vectors F which take values in $\mathbb{R}^N$. These vectors are indexed on the image grid

$$F = (F_k)_{k\in\mathscr{G}} \quad \text{where} \quad \mathscr{G} = \{-n/2+1,\dots,n/2\}^2,$$

(for simplicity we assume that n is even) and we use periodic boundary conditions. Without loss of generality, we assume that they have zero mean $\mathbb{E}(F) = 0$. Such a random vector is thus entirely characterized by its (square root) power spectrum density (PSD)

$$\forall\, \omega \in \mathscr{G}, \quad P_F(\omega) = \mathbb{E}(|\hat{F}(\omega)|^2)^{1/2}$$

where we define the Fourier transform of a vector or a random vector as

$$\forall\, \omega \in \mathscr{G}, \quad \hat{F}(\omega) = \sum_{k\in\mathscr{G}} F_k e^{\frac{2i\pi}{n}\langle k,\omega\rangle}$$

$$\text{where} \quad \langle k, \omega\rangle = k_1\omega_1 + k_2\omega_2.$$

We remind that once the power-spectrum P_F of F is known, F is recovered by

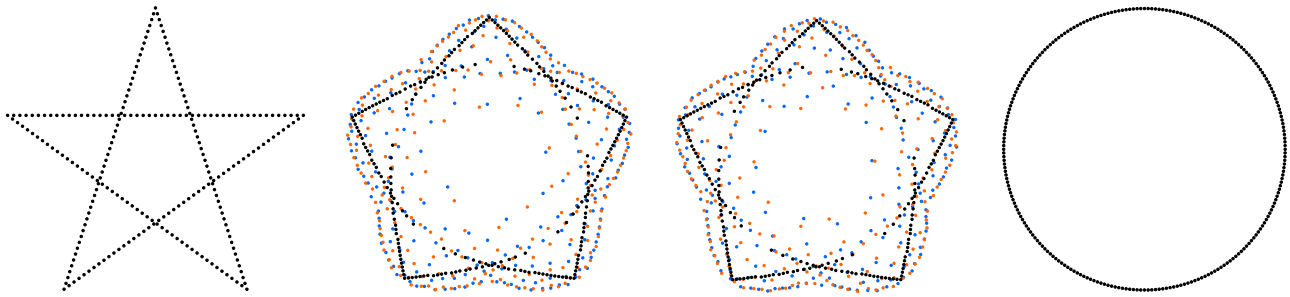
$$\hat{F}(\omega) = P_F(\omega)\cdot\hat{W}(\omega) \quad \text{where} \quad W \sim \mathscr{N}(0, \mathrm{Id}_N). \tag{46}$$

It is thus easy to draw a realization f of the vector F by convolving the inverse Fourier transform of P_F (the so-called texton, see [13]) by a realization w of the white noise W, i.e., computing $\hat{f} = P_F \cdot \hat{w}$, where $\cdot$ denotes entry-wise multiplication.

In this spot noise model, it is customary (see [16]) to learn the input Gaussian models $\{F^{[i]}\}_{i\in I}$ by estimating their PSD with a maximum likelihood estimation, which corresponds to estimating the covariance using the empirical periodogram

$$\forall\, i \in I, \quad \forall\, \omega \in \mathscr{G}, \quad P_{F^{[i]}}(\omega) = |\hat{f}^{[i]}(\omega)|.$$

We also use this estimation, which, despite its simplicity, gives good visual performances, see [15].



(a) μ_0 (b) $\mu^S_{1/2}$ (uniform Θ) (c) $\mu^S_{1/2}$ (random Θ) (d) μ_1

Fig. 4 Influence of the initialization $X^{[0]}$ and the directions set Θ (here $|\Theta| = 10^3$) on our Lagrangian discretization of the sliced barycenter. The black point cloud corresponds to the Wasserstein interpolation $\mu^W_{1/2}$ of the two distributions μ_0 and μ_1. The red and blue point clouds correspond to the sliced Wasserstein barycenters obtained with different settings: (b) using two random point clouds initializations for $X^{[0]}$ with the same set of directions Θ (equi-spaced on the circle); (c) using the same initializations $\mu_{X^{[0]}} = \mu_0$ but with different uniformly sampled random directions Θ.

$t = 0$ $t = 1/4$ $t = 1/2$ $t = 3/4$ $t = 1$

Fig. 5 The Radon geodesic $\mu_t = \mathrm{Bar}^R_{\mathbb{R}^d}((\mu_0,\mu_1),(t,1-t))$ between two anisotropic Gaussians is not Gaussian.

Optimal transport barycenter of SN models. We introduce a texture mixing method that performs the interpolation of the PSD using optimal transport. The rational of this method is to operate the mixing with geometric warpings of the spectral modes of the textures. The method is thus adapted to deal with micro-textures which exhibit a high degree of sparsity in the Fourier domain, i.e., which PSD are composed of a few localized spikes. This class of sparse spectral textures has been shown in [17] to be a powerful way to approximate more complicated textures for procedural texture synthesis.

We define the measure associated to the PSD of the Gaussian model $F^{[i]}$

$$\forall i \in I, \quad \mu_i = \frac{1}{\sum_{\omega\in\mathscr{G}} P_{F^{[i]}}(\omega)} \sum_{\omega\in\mathscr{G}} P_{F^{[i]}}(\omega)\delta_\omega \in \mathscr{M}_1^+(\mathbb{R}^2).$$

The barycenter measure is defined as

$$\forall \lambda \in \Lambda_I, \quad \mu^{(\lambda)} = \mathrm{Bar}^R_{\mathbb{R}^2}(\mu_i,\lambda_i)_{i\in I}.$$

Note that this measure exhibits central symmetry because of (11) and Proposition (8).

This barycenter measure is approximated using the Eulerian discretized Radon barycenter described in Section 3.4, to obtain a resulting measure

$$\bar{\mu}^{(\lambda)} = \mathrm{Bar}^R_{\mathscr{G}}(\mu_i,\lambda_i)_{i\in I}.$$

By construction of this algorithm, this measure is supported on the grid $\mathscr{G}$ and also exhibits central symmetry. It can thus be written as

$$\bar{\mu}^{(\lambda)} = \sum_{\omega\in\mathscr{G}} P_{F_\lambda}(\omega)\delta_\omega.$$

This thus defines a stationary Gaussian random vector F_λ through its PSD P_{F_λ}. This Gaussian vector is our interpolated model, which can be synthesized following (46).

Examples. We demonstrate our Radon barycenter of power spectrum densities on several examples. A sparse hand-designed power spectrum is interpolated in Fig. 10 and a more natural, less sparse, power spectrum is used in Fig. 11. We handle colors by convolving the interpolated power spectrum of each color channel by the same white noise. Although the decoupling of color channels could occasionally lead to color artifacts, we did not observe such effects on our set of examples (further examples can be see in the additional material). We hence leave the investigation of perceptually decoupled color spaces or the joint transportation of color channels for future work.

Comparison with linear interpolation. In [15], the authors also use optimal transport to perform SN model interpolation. Their approach is however radically different since they compute optimal transport geodesics in the space of Gaussian distributions in $\mathbb{R}^N$, which has a closed form solution. In contrast, we propose to compute the transportation of PSD in $\mathbb{R}^2$, viewed as discrete distributions of N Diracs. For grayscale textures, the method detailed in [15] thus boils down to a linear interpolation of the PSD, i.e., they define the PSD of the barycentric model $\tilde{F}_\lambda$ as

$$\forall \lambda \in \Lambda_I, \quad P_{\tilde{F}_\lambda} = \sum_{i\in I} \lambda_i P_{F^{[i]}}. \tag{47}$$

The effect achieved by our Radon barycenter differs from [15]. As shown in Figure 10 and 11, we believe our method is geometrically more meaningful when dealing with textures that have a sparse Fourier expansion, while [15] deal with denser spectra more appropriately. Sparse spectra can occur, for instance, for textures with approximately periodic tiling of repetitive patterns.

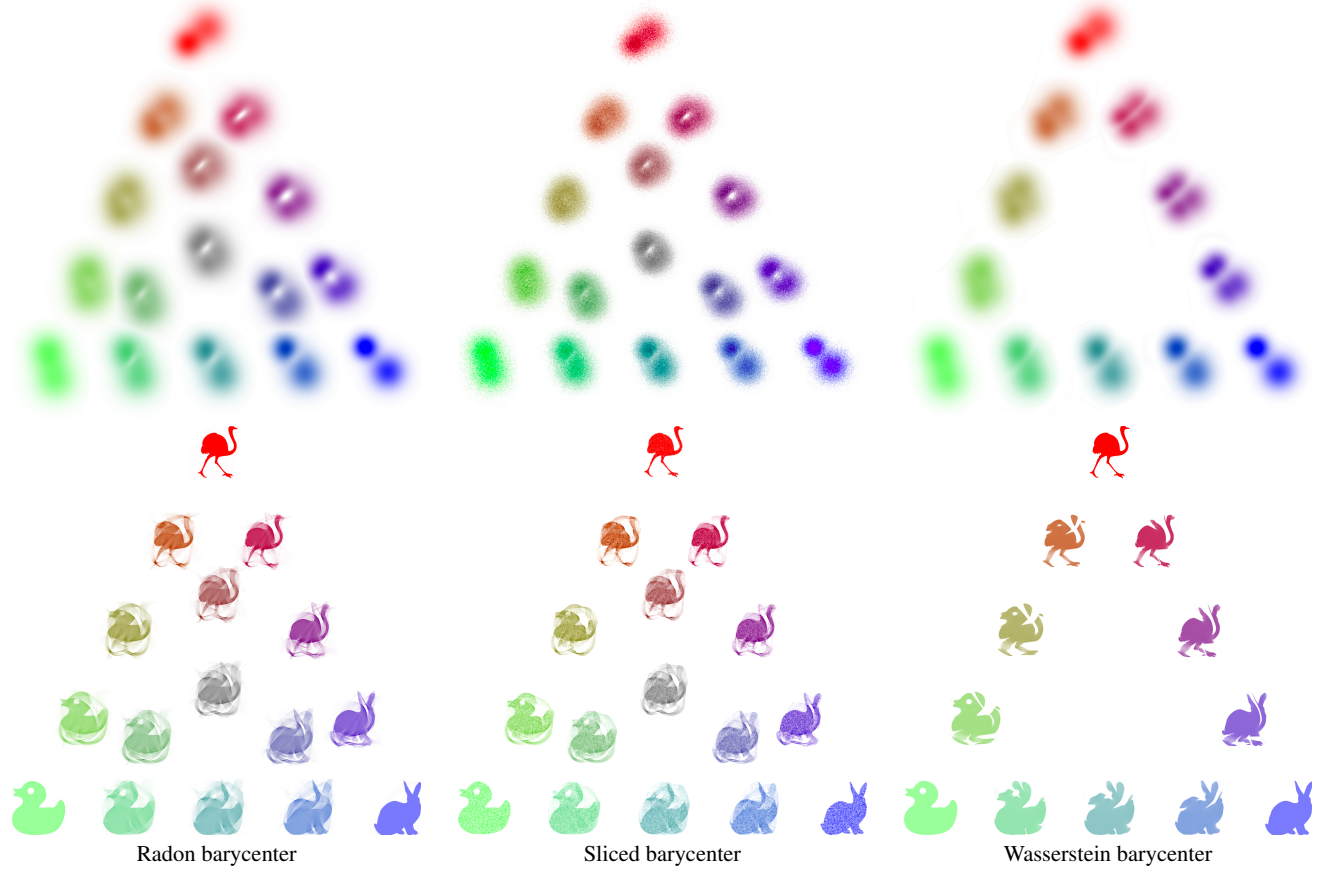


Fig. 6 Comparison of $\mathrm{Bar}^R_{\mathbb{R}^d}$, $\mathrm{Bar}^S_{\mathbb{R}^d}$ and $\mathrm{Bar}^W_{\mathbb{R}^d}$ (computed using the method detailed in [25]).

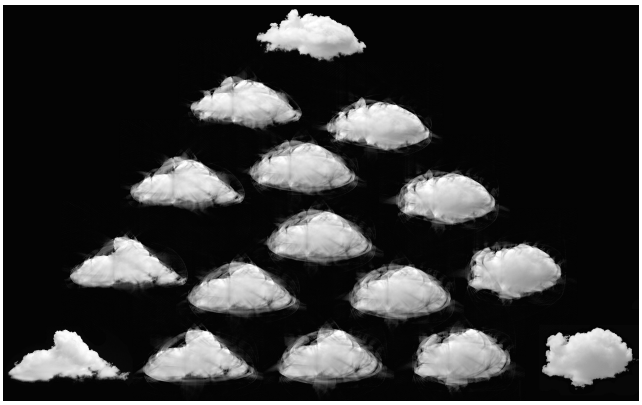


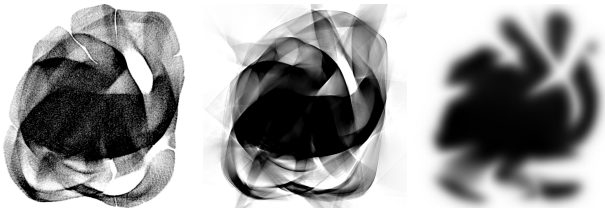
Fig. 7 Image warping using the Radon barycenter exhibits artifacts.

Fig. 8 Top: Radon barycenter $\mathrm{Bar}^R_{\mathbb{R}^d}$ of four 2-D distributions with equal weights. Bottom : Same experiment with SW2, using $N = 4\,10^4$ points samples, $|\Theta| = 100$ directions and a gaussian kernel with standard deviation $\sigma = 20/512$ to estimate the corresponding densities. .

5.5 Application to Color Palette Manipulation

In this section we investigate the benefit of our Sliced Wasserstein barycenter for two applications: harmonizing colors in an image sequence, and grading colors of a single image. Color harmonization is the process of bringing the colors of input images to an average color distribution such that the images end up looking more similar. This has several applications such as, for instance, image stitching or enforcing temporal coherence of colors in movies. The second application allows for the editing of a single image by bringing its colors closer to a set of photographs exhibiting particular color palettes. This process is called color grading, and finds applications in photograph enhancement.

Lagrangian color palette. We consider a color image represented as a vector $X \in \mathbb{R}^{N\times 3}$ of N pixels, so that $X = (X_k)_{k=1,\dots,N}$ where each pixel $X_k \in \mathbb{R}^3$ stores the value of a pixel indexed by k. In the following, we use the YCbCr color space because of its ability to decorrelate color channels, although other color spaces may be used (e.g., the CIE-Lab



$\mathrm{Bar}^S_{\mathbb{R}^d}$ $\mathrm{Bar}^R_{\mathbb{R}^d}$ Cuturi et al.

Fig. 9 Comparison of three methods (Sliced, Radon, and the one presented in [10]) to compute isobarycenters (i.e. using $\lambda = (1,1,1)/3$) of the three input densities displayed at the vertices of Figure 6, bottom.

advocated in [29]). The color distribution of this image is a measure μ_X defined in $\mathbb{R}^3$, and describes the color palette. We naturally represent this color distribution using a Lagrangian discretization, as defined in (36), by essentially storing pixel colors as a point cloud in the space of colors. Note that the Lagrangian discretization (36) defining μ_X is automatically normalized so that $\mu_X \in \mathscr{M}_1^+(\mathbb{R}^2)$. We hence compute the average distribution of multiple images distributions using our (Lagrangian) Sliced Wasserstein Barycenter detailed in Section 4.3.

Color palette transfer. Before detailing our main application to color palette barycenters, we illustrate our stochastic gradient descent (Section 4.4). This descent allows for the computation of an approximate Sliced transport map T^S between the color palette $\mu_{X^{[0]}}$ of an input image $X^{[0]}$ and the model palette μ_Y of an image Y, where $X^{[0]}, Y \in \mathbb{R}^{N\times 3}$. The resulting image $X^\star$ is obtained as the limit of the stochastic gradient descent steps (42) until convergence

$$X^{[\ell]} \overset{\ell\to+\infty}{\longrightarrow} X^\star, \tag{48}$$

as described in (43).

We illustrate our technique in Fig. 12. This process generalizes the algorithm introduced in [26] that uses $|\Theta_\ell| = 3$ orthogonal directions at each step. While we make use of an exact Lagrangian method by sorting pixel values, Pitié et al. discretize histograms and use the cumulative histogram and pseudo-inverse approach (Eqs. 16 and 17). The lower complexity of [26] comes at the expense of a discretization which can lead to quantization errors and limits convergence.

Color palette barycenter. We consider a set $\{X^{(i)}\}_{i\in I}$ of color images, as well as a particular input color image $X^{[0]}$. Using (37), we define the color palette $\mu_{X^\star}$, the barycenter of the input palettes $\mu_{X^{(i)}}$, as the Sliced Wasserstein barycenter $\mu_{X^\star} \approx \mathrm{Bar}^S_{\mathbb{R}^d}\left(\mu_{X^{(i)}}, \lambda_i\right)_{i\in I}$ with weights $\lambda \in \Lambda_I$.

Color image harmonization and color grading. In order to adjust colors in an image, we are interested in an image $X^\star$ visually similar to $X^{[0]}$, but whose palette closely matches the palette barycentre $\mu_{X^\star}$. Similarly to the simple color transfer application (see (48)), we obtain this image by performing the gradient descent iterations (39) with initialization $X^{[0]}$, and define $X^\star$ as the limit image $X^{[\ell]} \overset{\ell\to+\infty}{\longrightarrow} X^\star$.

However, highly non-linear color transformations can create undesirable visual artifacts. We therefore use an iterative post-processing technique introduced in [27] to regularize the transportation map $X_k^{[0]} \mapsto X_k^\star$. We refer the interested reader to [27] for further details.

For color harmonization, we apply this process successively to each image in an input sequence $\{X^{(i)}\}_{i\in I}$, by initializing $X^{[0]}$ with $X^{(i)}$ for each i. For color grading, we instead apply the palette barycenter to an arbitrary input image $X^{[0]}$.

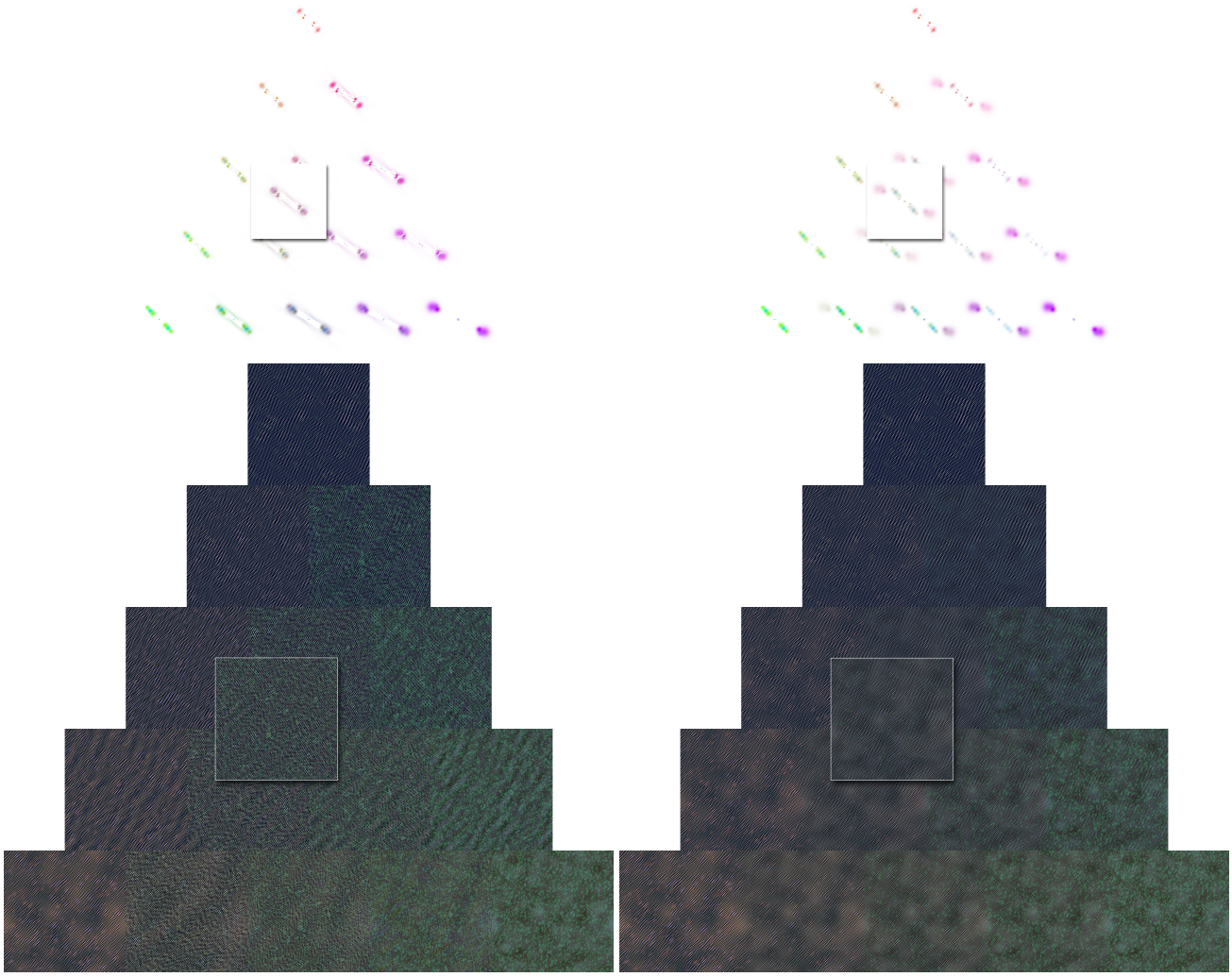
Table 1 Coordinates w used to define the weights $\lambda = w/(\sum_i w_i)$ for the color transfer in Figure 13.

		$(0,0,1)$		
	$(1,0,3)$		$(0,1,3)$	
$(1,0,1)$		$(1,1,2)$		$(0,1,1)$
	$(2,1,1)$		$(1,2,1)$	
$(3,0,1)$		$(1,1,0)$		$(0,3,1)$
	$(3,1,0)$		$(1,3,0)$	
$(1,0,0)$				$(0,1,0)$

Examples. Figure 14 shows an example of harmonization, where the color palette is defined as the iso-barycenter of three input color palettes. In Figure 13, the image $X^{[0]}$ to be modified is not contained in the set of input pictures $\{X^{(i)}\}_{i\in I}$. This allows for the user to navigate over the simplex of color palettes to select the desired one. Table 1 provides the corresponding weights for Figure 13.

6 Conclusion

We introduce two novel different definitions of barycenters of multi-dimensional measures based on one-dimensional optimal transport. We show that these Radon and Sliced Wasserstein Barycenters enjoy the same invariance properties as the usual Wasserstein barycenter. They both minimize variational problems, which are almost identical, up to the lack of surjectivity of the Radon transform. We estimate this deviation to be negligible on a set of examples. We introduce Lagrangian and Eulerian discretization schemes, which enable the approximation of these barycenters with fast algorithms. The computational time is orders of magnitude faster than the Wasserstein barycenter counterpart for two input measures. Furthermore, they can be applied to more than two input densities. We show on several numerical examples that, while these barycenters exhibit significant geometrical differences with respect to the Wasserstein barycenter, they appear to be very well suited to several applications in image processing and computer graphics.



(a) Radon barycenter (our approach) (b) Linear interpolation [15]

Fig. 10 (a) Eulerian Radon barycenter interpolates sparse amplitude spectra. (b) linear interpolation of the amplitude spectrum (47), as performed in [15]. The top row shows the interpolated spectra P_{F_λ}.

Acknowledgment

We thank Marco Cuturi for applying his method to our dataset and for sharing his results. We thank Thouis R. Jones for useful feedback on our draft, and anonymous reviewers for their help in improving this paper. We also thank the authors of all the images used to demonstrate our color transfers. This work has been partially supported by NSF CGV-1111415. Gabriel Peyré acknowledges support from the European Research Council (ERC project SIGMA-Vision).

A Proofs of Section 2

Proof of Proposition 1. From the definition (5), one verifies that

$$\mathrm{W}_{\mathbb{R}^d}(\varphi_{s,u}\sharp\mu_1,\varphi_{s,u}\sharp\mu_2) = s\mathrm{W}_{\mathbb{R}^d}(\mu_1,\mu_2). \tag{49}$$

so that

$$\begin{aligned}\mathscr{E}_{s,u}(\mu) &= \sum_{i\in I}\lambda_i \mathrm{W}_{\mathbb{R}^d}(\varphi_{s,u}\sharp\mu_i,\mu)^2 \\ &= s^2\sum_{i\in I}\lambda_i \mathrm{W}_{\mathbb{R}^d}(\mu_i,\varphi_{s,u}^{-1}\sharp\mu)^2 = s^2\mathscr{E}_{1,0}(\tilde{\mu}).\end{aligned}$$

where we have introduced the following change of variable

$$\mu = \varphi_{s,u}\sharp\tilde{\mu} \quad\Longleftrightarrow\quad \tilde{\mu} = \varphi_{s,u}^{-1}\sharp\mu,$$

(note that $\varphi_{s,u}^{-1} = \varphi_{s^{-1},-s^{-1}u}$). One thus has

$$\operatorname*{argmin}_{\mu}\mathscr{E}_{s,u}(\mu) = \varphi_{s,u}\sharp\operatorname*{argmin}_{\tilde{\mu}}\mathscr{E}_{1,0}(\tilde{\mu})$$

which proves (7). Property (8) is proved similarly. Properties (9) and (11) directly follow from (8). □

Proof of Proposition 2. We aim at determining $(s^\star,u^\star)$ such that

$$\mu^\star \in \mathrm{Bar}^W_{\mathbb{R}^d}(\mu_i,\lambda_i)_{i\in I} \quad \text{where} \quad \begin{cases}\mu^\star = \varphi^\star\sharp\mu,\\ \mu_i = \varphi_i\sharp\mu,\end{cases}$$

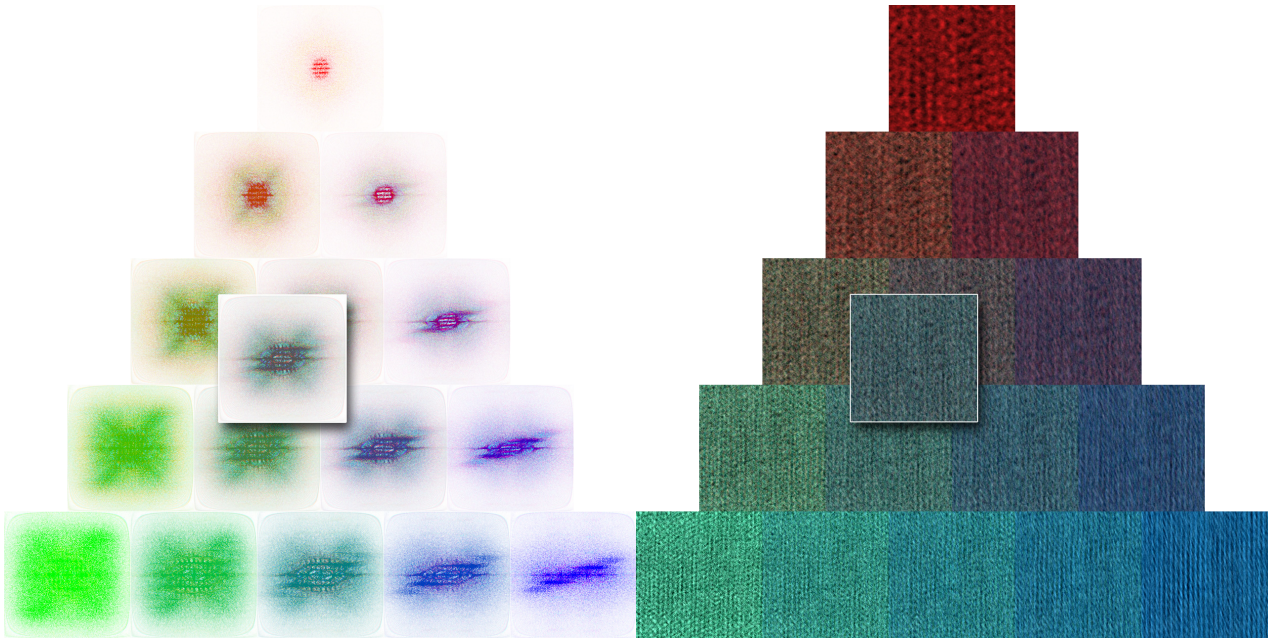


Fig. 11 Eulerian Radon barycenter applied to the mixing of natural textures.

(a) input image $X^{[0]}$

(b) input model Y

(c) our result $X^\star$

Fig. 12 Our stochastic gradient descent (c) can be used to transfer the colors of a model image (b) to an input image (a). We generalize the method of Pitié et al. [26] as described in Sec. 5.5

and where for simplicity we have denoted $\varphi_i = \varphi_{s_i,u_i}$ and $\varphi^\star = \varphi_{s^\star,u^\star}$. First, let us notice that

$$\varphi_{s,u}(x) = \nabla\left(\frac{s}{2}\|x+u/s\|^2\right),$$

so that the set $\mathscr{T}$ of maps of the form $\varphi_{s,u}$ is a subset of gradients of convex functions. This point is important since optimal maps between μ_i and $\mu^\star$ are characterized as the gradient of convex functions that push forward μ_i onto $\mu^\star$, see [32]. Following [1], we thus only need to show that

$$\sum_{i\in I}\lambda_i T_i = \mathrm{Id}_{\mathbb{R}^d} \quad \text{where} \quad T_i = \varphi^\star\circ\varphi_i^{-1} = \varphi_{\tilde{s}_i,\tilde{u}_i}$$

$$\text{where} \quad \begin{cases} \tilde{s}_i = s^\star s_i^{-1} \\ \tilde{u}_i = u^\star - s^\star s_i^{-1} u_i \end{cases}$$

since $T_i\sharp\mu_i = \mu^\star$ and $T_i \in \mathscr{T}$ is a gradient of a convex function. So that $\mu^\star$ is a barycenter if and only if

$$\sum_{i\in I}\lambda_i T_i = \sum_{i\in I}\lambda_i\varphi_{\tilde{s}_i,\tilde{u}_i}$$
$$= \varphi_{\sum_{i\in I}\lambda_i\tilde{s}_i,\sum_{i\in I}\lambda_i\tilde{u}_i} = \mathrm{Id}_{\mathbb{R}^d} = \varphi_{1,0}.$$

This in turn is equivalent to the relationships

$$\sum_{i\in I}\lambda_i\tilde{s}_i = 1 \quad \text{and} \quad \sum_{i\in I}\lambda_i\tilde{u}_i = 0,$$

which corresponds to (14). □

Proof of Proposition 3. The proof is done in [1] for $\mu = \mu_j$ for some $j \in I$, which is supposed to be absolutely continuous. It extends to an arbitrary measure μ. □

Proof of Corollary 1. When using μ, the uniform and normalized measure on $[0,1]$, with the notation of Proposition (3), one has $T_i = C_{\mu_i}^+$. This is indeed a classical result for 1-D optimal transport, see for instance [1], Section 6.1. One then recognizes that formula (18) is the same as formula (15). □

Proof of Proposition 4. One has

$$\nu^\star \in \operatorname*{argmin}_{\nu\in\mathscr{M}_1^+(\Omega^d)}\sum_{i\in I}\lambda_i \mathrm{W}_{\Omega^d}(\nu_i,\nu)^2$$
$$= \operatorname*{argmin}_{\nu\in\mathscr{M}_1^+(\Omega^d)}\int_{\mathbb{S}^{d-1}}\sum_{i\in I}\lambda_i \mathrm{W}_{\mathbb{R}}(\nu_i^\theta,\nu^\theta)^2 \mathrm{d}\theta.$$

This is equivalent to the fact that for almost all $\theta\in\mathbb{S}^{d-1}$, one has

$$\nu^{\star,\theta} \in \mathrm{Bar}_{\mathbb{R}}^W(\nu_i^\theta,\lambda_i)_{i\in I}.$$

□

Proof of Proposition 5. Proof of (20). Similarly to the proof of (7), the proof of (20) is obtained by using the following invariance of the Wasserstein distance on Ω^d

$$\mathrm{W}_{\Omega^d}(\psi_{s,u}\sharp\nu_1,\psi_{s,u}\sharp\nu_2) = s\mathrm{W}_{\Omega^d}(\nu_1,\nu_2). \tag{50}$$

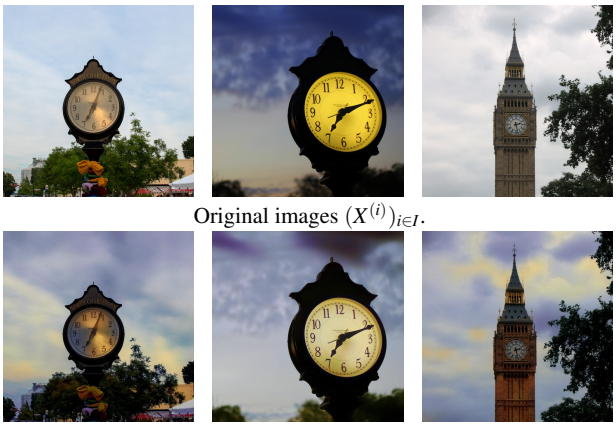
Fig. 13 Color manipulation by transferring the colors of $|I| = 3$ photographs $\{X_i\}_{i\in I}$ (shown at the vertices of the triangle, right) to the initial photograph $X^{[0]}$ (left) to obtain $X^\star$ which varies in the triangle as a function of the convex weights $\lambda \in \Lambda_I$. Additional results can be seen in supplemental material.

Proof of (21). One has that $\nu^\star \in \mathrm{Bar}^W_{\Omega^d}(\psi_{s_i,u_i}\sharp\nu, \lambda_i)_{i\in I}$ is equivalent to which gives the desired result. □

for almost all $\theta \in \mathbb{S}^{d-1}$, $\quad (\nu^\star)^\theta \in \mathrm{Bar}^W_{\mathbb{R}}(\varphi_{s_i,\langle u_i,\theta\rangle}\sharp\nu^\theta, \lambda_i)_{i\in I}$.

Using the property of proposition 2 for $d = 1$, one obtains that

$\mathrm{Bar}^W_{\mathbb{R}}(\varphi_{s_i,\langle u_i,\theta\rangle}\sharp\nu^\theta, \lambda_i)_{i\in I} \ni \varphi_{s^\star,\langle u^\star,\theta\rangle}\sharp\nu^\theta$,



Original images $(X^{(i)})_{i\in I}$.

Harmonized images $\{X^{(i,\star)}\}_{i\in I}$.

Fig. 14 Color harmonization of an image sequence, using $\lambda_i = 1/|I|$ to compute the iso-barycenter (here $|I| = 3$).

B Proofs of Section 3

Proof of Proposition 6. For all $g \in \mathscr{C}_0(\Omega^d)$, one has

$$\begin{aligned}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}} g(t,\theta)\mathrm{d}(R(\mu)^\theta)(t)\mathrm{d}\theta &= \int_{\Omega^d} g(t,\theta)\mathrm{d}(R(\mu))(t,\theta)\\ &= \int_{\mathbb{R}^d}(R^*g)(x)\mathrm{d}\mu(x)\\ &= \int_{\mathbb{R}^d}\int_{\mathbb{S}^{d-1}} g(P_\theta(x),\theta)\mathrm{d}\theta\mathrm{d}\mu(x)\\ &= \int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}} g(y,\theta)\mathrm{d}(P_\theta\sharp\mu)(y)\mathrm{d}\theta.\end{aligned}$$

□

Proof of Lemma 1. *Proof of* (26)*:* For all $g \in \mathscr{C}_0(\Omega^d)$, one has

$$\begin{aligned}\int_{\mathbb{R}^d} g\mathrm{d}[R(\varphi_{s,u}\sharp\mu)] &= \int_{\mathbb{R}^d} R^*(g)\mathrm{d}[\varphi_{s,u}\sharp\mu]\\ &= \int_{\mathbb{R}^d}\int_{\mathbb{S}^{d-1}} g(\langle sx+u,\theta\rangle,\theta)\mathrm{d}\theta\mathrm{d}\mu(x)\\ &= \int_{\mathbb{R}^d}\int_{\mathbb{S}^{d-1}} (g\circ\psi_{s,u})(\langle x,\theta\rangle,\theta)\mathrm{d}\theta\mathrm{d}\mu(x)\\ &= \int_{\mathbb{R}^d}(g\circ\psi_{s,u})\mathrm{d}[R(\mu)]\\ &= \int_{\mathbb{R}^d} g\mathrm{d}[\psi_{s,u}\sharp R(\mu)]\end{aligned}$$

Proof of (27): First we notice, using (22), that

$$\begin{aligned}R(f\circ\varphi_{s,u})(t,\theta) &= \int_{\mathbb{R}^{d-1}} f\left(s(t\theta+U_\theta\gamma)+u\right)\mathrm{d}\gamma\\ &= \int_{\mathbb{R}^{d-1}} f\left(st\theta+U_\theta s\gamma+\langle u,\theta\rangle\theta+U_\theta(U_\theta)^T u\right)\mathrm{d}\gamma\\ &= \int_{\mathbb{R}^{d-1}} f\left((st+\langle u,\theta\rangle)\theta+U_\theta(s\gamma+(U_\theta)^T u)\right)\mathrm{d}\gamma\\ &= s^{1-d}\int_{\mathbb{R}^{d-1}} f\left(\psi_{s,u}(t,\theta)\theta+U_\theta\gamma'\right)\mathrm{d}\gamma'\end{aligned}$$

which proves

$$R(f\circ\varphi_{s,u}) = s^{1-d}R(f)\circ\psi_{s,u} \tag{51}$$

We write $H = (R^*R)^{-1}$ the filtering operator with kernel h^+. One has, for smooth functions $f \in \mathscr{S}(\mathbb{R}^d)$, denoting $\mathscr{F}(f) = \hat{f}$,

$$\begin{aligned}\mathscr{F}(H(f\circ\varphi_{s,u})) &= c^{-1}\|\omega\|^{1-d}\hat{f}(s\omega)e^{-\mathrm{i}\langle\omega,u\rangle},\\ \mathscr{F}(H(f)\circ\varphi_{s,u}) &= c^{-1}\|s\omega\|^{1-d}\hat{f}(s\omega)e^{-\mathrm{i}\langle\omega,u\rangle},\end{aligned}$$

and hence

$$H(f)\circ\varphi_{s,u} = s^{1-d}H(f\circ\varphi_{s,u}). \tag{52}$$

This shows, using (51) and (52) that for all $f \in \mathscr{D}(\mathbb{R}^d)$,

$$\begin{aligned}\int_{\mathbb{R}^d} f\mathrm{d}[R^+(\psi_{s,u}\sharp\nu)] &= \int_{\mathbb{R}^d}(RHf)\circ\psi_{s,u}\mathrm{d}\nu\\ &= s^{d-1}\int_{\mathbb{R}^d} R(H(f)\circ\varphi_{s,u})\mathrm{d}\nu\\ &= \int_{\mathbb{R}^d} RH(f\circ\varphi_{s,u})\mathrm{d}\nu\\ &= \int_{\mathbb{R}^d} f\mathrm{d}[\varphi_{s,u}\sharp R^+(\nu)]\end{aligned}$$

Proof of (28): the proof is similar to the one of (26). □

Proof of Proposition 8. Using Lemma 1, one has

$$\begin{aligned}\mathrm{Bar}^R_{\mathbb{R}^d}(\varphi_{s,u}\sharp\mu_i,\lambda_i)_{i\in I} &= R^+\mathrm{Bar}^W_{\Omega^d}(R(\varphi_{s,u}\sharp\mu_i),\lambda_i)_{i\in I}\\ &= R^+\mathrm{Bar}^W_{\Omega^d}(\psi_{s,u}\sharp(R(\mu_i)),\lambda_i)_{i\in I}\\ &= R^+\psi_{s,u}\sharp\mathrm{Bar}^W_{\Omega^d}(R(\mu_i),\lambda_i)_{i\in I}\\ &= \varphi_{s,u}\sharp R^+\mathrm{Bar}^W_{\Omega^d}(R(\mu_i),\lambda_i)_{i\in I}\\ &= \varphi_{s,u}\sharp\mathrm{Bar}^R_{\mathbb{R}^d}(\mu_i,\lambda_i)_{i\in I}.\end{aligned}$$

which proves (7) for $\mathrm{Bar}^R_{\mathbb{R}^d}$. Property (8) for $\mathrm{Bar}^R_{\mathbb{R}^d}$ is proved similarly using (28). □

Proof of Proposition 9. One has

$$\begin{aligned}\mathrm{Bar}^R_{\mathbb{R}^d}(\varphi_{s_i,u_i}\sharp\mu,\lambda_i)_{i\in I} &= R^+\mathrm{Bar}^W_{\Omega^d}(R(\varphi_{s_i,u_i}\sharp\mu),\lambda_i)_{i\in I}\\ &= R^+\mathrm{Bar}^W_{\Omega^d}(\psi_{s_i,u_i}\sharp R(\mu),\lambda_i)_{i\in I}\\ &= R^+\psi_{s^\star,u^\star}\sharp\mathrm{Bar}^W_{\Omega^d}(R(\mu),\lambda_i)_{i\in I}\\ &= \varphi_{s^\star,u^\star}\sharp R^+\mathrm{Bar}^W_{\Omega^d}(R(\mu),\lambda_i)_{i\in I}\\ &= \varphi_{s^\star,u^\star}\sharp\mathrm{Bar}^R_{\mathbb{R}^d}(\mu,\lambda_i)_{i\in I},\end{aligned}$$

which proves (13) for $\mathrm{Bar}^R_{\mathbb{R}^d}$. □

C Proof of Section 4

Proof of Proposition 10. Property (34) is a re-statement of property (19). Property (35) corresponds to the change of variable $\nu = R\mu \in \mathrm{Im}(R)$ in (32), which is a bijection thanks to the injectivity of R, see proposition 7. □

Proof of Proposition 11. The proof is the same as Proposition 1, replacing the invariance (49) by

$$\begin{aligned}\mathrm{SW}_{\mathbb{R}^d}(\varphi_{s,u}\sharp\mu_1,\varphi_{s,u}\sharp\mu_2) &= \mathrm{W}_{\Omega^d}(R(\varphi_{s,u}\sharp\mu_1),R(\varphi_{s,u}\sharp\mu_2))\\ &= \mathrm{W}_{\Omega^d}(\psi_{s,u}\sharp R(\mu_1),\psi_{s,u}\sharp R(\mu_1))\\ &= \mathrm{W}_{\Omega^d}(R(\mu_1),R(\mu_1))\\ &= \mathrm{SW}_{\mathbb{R}^d}(\mu_1,\mu_2),\end{aligned}$$

where we have used the invariance (50) of the Wasserstein distance on Ω^d. □

Proof of Proposition 12. One has,

$$\forall\,\theta\in\mathbb{S}^{d-1},\quad P_\theta\sharp\varphi_{s,u}\sharp\mu = \varphi_{s,\langle u,\theta\rangle}\sharp P_\theta\sharp\mu.$$

Thus, for an arbitrary $\tilde{\mu} \in \mathscr{M}_1^+(\mathbb{R}^d)$, one has

$$\begin{aligned}
&\sum_{i\in I} \lambda_i \mathrm{W}_{\mathbb{R}}(P_\theta \sharp(\varphi_{s_i,u_i}\sharp\mu), P_\theta\sharp\tilde{\mu})^2 \\
&\quad = \sum_{i\in I} \lambda_i \mathrm{W}_{\mathbb{R}}(\varphi_{s_i,\langle u_i,\theta\rangle}\sharp(P_\theta\sharp\mu), P_\theta\sharp\tilde{\mu})^2 \\
&\quad \geqslant \sum_{i\in I} \lambda_i \mathrm{W}_{\mathbb{R}}(\varphi_{s_i,\langle u_i,\theta\rangle}\sharp(P_\theta\sharp\mu), \varphi_{s^\star,\langle u^\star,\theta\rangle}\sharp(P_\theta\sharp\mu))^2 \\
&\quad = \sum_{i\in I} \lambda_i \mathrm{W}_{\mathbb{R}}(P_\theta\sharp(\varphi_{s_i,u_i}\sharp\mu), P_\theta\sharp(\varphi_{s^\star,u^\star}\sharp\mu))^2
\end{aligned}$$

where the inequality comes from the properties of 1-D Wasserstein barycenters. Integrating the resulting inequality with respect to $\theta \in \mathbb{S}^{d-1}$ gives

$$\sum_i \lambda_i \mathrm{SW}_{\mathbb{R}^d}(\varphi_{s_i,u_i}\sharp\mu, \tilde{\mu})^2 \geqslant \sum_i \lambda_i \mathrm{SW}_{\mathbb{R}^d}(\varphi_{s_i,u_i}\sharp\mu, \varphi_{s^\star,u^\star}\sharp\mu)^2.$$

This inequality is an equality if and only for almost all $\theta \in \mathbb{S}^{d-1}$, one has

$$P_\theta\sharp\tilde{\mu} = P_\theta\sharp(\varphi_{s^\star,u^\star}\sharp\mu)$$

so that, using Proposition (7), this corresponds to $\tilde{\mu} = \varphi_{s^\star,u^\star}\sharp\mu$. Since the measure $\tilde{\mu}$ is arbitrary, this gives the desired result. This proves (13) in the case $\mathrm{Bar}^S_{\mathbb{R}^d}$. □

D Proof of Theorem 1

Notations. Without loss of generality, for a fixed $Y \in \mathbb{R}^{d\times N}$, we study the smoothness of

$$\forall X \in \mathbb{R}^{d\times N}, \quad \mathscr{E}(X) = \frac{1}{2}\mathrm{SW}_{\mathbb{R}^d}(\mu_X,\mu_Y)^2 = \int_{\mathbb{S}^{d-1}} \mathscr{E}_\theta(X)\mathrm{d}x$$

$$\text{where} \quad \mathscr{E}_\theta(X) = \frac{1}{2}\mathscr{W}(X_\theta, Y_\theta)^2.$$

We have used, for $x,y \in \mathbb{R}^N$, the shorthand notation

$$\mathscr{W}(x,y) = \mathrm{W}_{\mathbb{R}}(\mu_x,\mu_y).$$

The result of Theorem 1 then follows by summations of such functionals.

We define $\mathbb{U}(N,d)$ to be vectors of $\mathbb{R}^{d\times N}$ with distinct entries:

$$\mathbb{U}(N,d) = \left\{ W=(W_1,\ldots,W_N) \in \mathbb{R}^{d\times N} \,;\, \forall i\neq j, X_i \neq X_j \right\}.$$

The hypothesis is that $X \in \mathbb{U}(N,d)$. One has

$$\mathscr{E}_\theta(X) = \frac{1}{2}\|X_\theta - Y_\theta\circ\sigma_\theta\|^2 \quad \text{where} \quad \sigma_\theta = \sigma_X^\theta \circ (\sigma_Y^\theta)^{-1}$$

is a permutation depending on both X and Y. Note that the permutation involved are not necessarily unique, and are assumed to be arbitrary valid sorting permutations.

For $X \in \mathbb{R}^{N\times d}$ and $\varepsilon > 0$ we introduce

$$\Theta_\varepsilon(X) = \left\{ \theta \in \mathbb{S}^{d-1} \,;\, \forall \|\delta\|_{\mathbb{R}^{N\times d}} \leqslant \varepsilon, \quad X_\theta + \delta_\theta \in \mathbb{U}(N,1) \right\}.$$

This is the set of directions for which any perturbation of X of amplitude smaller than ε has a projection with disjoint points.

Overview of the proof. In the following, we thus aim at proving that $\mathscr{E}$ is C^1, that

$$\tilde{\nabla}\mathscr{E}(X) = \int_{\mathbb{S}^{d-1}} \tilde{\nabla}\mathscr{E}_\theta(X)\mathrm{d}\theta \quad \text{where} \quad \tilde{\nabla}\mathscr{E}_\theta(X) = (X_\theta - Y_\theta\circ\sigma_\theta)\theta$$

is indeed equal to $\nabla\mathscr{E}(X)$, and that this gradient is Lipschitz continuous.

The general strategy of the proof is to split the integration between the directions $\theta \in \Theta_\varepsilon(X)$, for which we can locally assume that the permutations σ_θ are constant (see Lemma 2), which in turn defines a smooth quadratic energy, and the remaining directions in $\Theta_\varepsilon(X)^c$, which are shown to have a negligible contribution to the energy and to the derivative (see Lemma 3).

Preparatory results. The following lemma shows that if $\theta \in \Theta_\varepsilon(X)$ the permutations σ_X^θ are stable to small perturbations of X.

Lemma 2. *Let $X \in \mathbb{U}(N,d)$. For all $\theta \in \Theta_\varepsilon(X)$, for all δ with $\|\delta\|_{\mathbb{R}^{N\times d}} \leqslant \varepsilon$, the permutation $\sigma^\theta_{X+\delta}$ that sorts $(\langle X_i+\delta_i,\theta\rangle)_i$ is uniquely defined and satisfies $\sigma^\theta_{X+\delta} = \sigma^\theta_X$.*

Proof. If one has $\sigma^\theta_{X+\delta} \neq \sigma^\theta_X$, then necessarily there exists some $t \in [0,1]$ such that $\sigma^\theta_{X+t\delta}$ is not uniquely defined, which is equivalent to $X_\theta + t\delta_\theta$ not being in $\mathbb{U}(N,1)$. Since $\|t\delta\|_{\mathbb{R}^{N\times d}} \leqslant \varepsilon$, this shows that $\theta \notin \Theta_\varepsilon(X)$. □

In order to prove Theorem 1, we need the following lemma.

Lemma 3. *For $X \in \mathbb{U}(N,d)$, one has*

$$\mathrm{Vol}(\Theta_\varepsilon(X)^c) = \int_{\Theta_\varepsilon(X)^c} \mathrm{d}\theta = O(\varepsilon). \tag{53}$$

Proof. One has $X_\theta + \delta_\theta \notin \mathbb{U}(N,1)$ if and only there exists a pair of points $u = X_i + \delta_i$ and $v = X_j + \delta_j$ with $i \neq j$ such that

$$\theta \in A(u,v) \quad \text{where} \quad A(u,v) = \left\{ \xi \in \mathbb{S}^{d-1} \,;\, \langle \xi, u-v\rangle = 0 \right\}$$

Note that $A(u,v)$ is a great circle of the sphere $\mathbb{S}^{d-1}$.

One can thus covers $\Theta_\varepsilon(X)^c$ using the union of all such circles $A(u,v)$, which shows

$$\Theta_\varepsilon(X)^c \subset \bigcup_{i\neq j} A_\varepsilon(X_i,X_j) \quad \text{where} \quad A_\varepsilon(x,y) = \bigcup_{\substack{\|u-x\|\leqslant\varepsilon \\ \|v-y\|\leqslant\varepsilon}} A(u,v)$$

Note that the geodesic distance d on the sphere $\mathbb{S}^{d-1}$ between two circles is equal to the angle between the normal to the planes of the circles

$$d(A(u,v),A(x,y)) = \mathrm{Angle}(u-v,x-y) = \mathrm{Angle}(x-y+\varepsilon w, x-y)$$

where $\|w\| \leqslant 2$. As $\varepsilon \to 0$, after some computations, one has the following asymptotic decay of the angle

$$\mathrm{Angle}(x-y+\varepsilon w, x-y) = O(\varepsilon/\|x-y\|)$$

and thus $d(A(u,v),A(x,y)) \leqslant C\varepsilon$ for some constant C. This proves that $\forall u,v$, one has

$$\begin{cases} \|u-x\| \leqslant \varepsilon \\ \|v-y\| \leqslant \varepsilon \end{cases} \quad \Longrightarrow \quad A(u,v) \subset B_{C\varepsilon}(x,y)$$

for some constant $C > 0$, where

$$B_\varepsilon(x,y) = \left\{ \xi \in \mathbb{S}^{d-1} \,;\, d(\xi, A(x,y)) \leqslant \varepsilon \right\}$$

One thus has

$$A_\varepsilon(x,y) \subset B_{C\varepsilon}(x,y).$$

The volume of the spherical band $B_{C\varepsilon}(x,y)$ of width $C\varepsilon$ is proportional to ε, and thus $\mathrm{Vol}(A_\varepsilon(x,y)) = O(\varepsilon)$. Since $\Theta_\varepsilon(X)^c$ is a finite union of such sets, one obtains the result. □

Proof of continuity. For each θ, the function $\mathscr{E}_\theta$ is continuous as a minimum of continuous functions. The function $\mathscr{E}$ being an integral of $\mathscr{E}_\theta$ on a compact set $\mathbb{S}^{d-1}$, it is thus continuous.

Proof of differentiability. Let $\delta \in \mathbb{R}^{N\times d}$ and $\varepsilon = \|\delta\|_{\mathbb{R}^{N\times d}}$. The definition of the Wasserstein distance reads

$$\mathscr{W}((X+\delta)_\theta, Y_\theta)^2 = \|(X_\theta+\delta_\theta)\circ\sigma^\theta_{X+\delta} - Y_\theta\circ\sigma^\theta_Y\|^2.$$

For all $\theta \in \Theta_\varepsilon(X)$, thanks to Lemma 2, $\sigma^\theta_{X+\delta} = \sigma^\theta_X$. One can thus compute the variation of the 1-D Wasserstein distance with respect to δ as

$$\mathscr{W}((X+\delta)_\theta, Y_\theta)^2 = \|X_\theta+\delta_\theta - Y_\theta\circ\sigma_\theta\|^2 \tag{54}$$
$$= \mathscr{W}(X_\theta, Y_\theta)^2 + \langle\tilde{\nabla}\mathscr{E}_\theta(X), \delta\rangle_{\mathbb{R}^{N\times d}} + \|\delta_\theta\|^2. \tag{55}$$

Note that the fact that σ^θ_Y might not be uniquely defined has no impact on the value of (55). One thus has

$$\mathscr{E}(X+\delta) - \mathscr{E}(X) - \langle\tilde{\nabla}\mathscr{E}(X), \delta\rangle_{\mathbb{R}^{N\times d}} = A(\delta) + B(\delta) + O(\|\delta\|^2_{\mathbb{R}^{N\times d}})$$

where

$$A(\delta) = \int_{\Theta_\varepsilon(X)^c} \left(\mathscr{W}(X_\theta+\delta_\theta, Y_\theta)^2 - \mathscr{W}(X_\theta, Y_\theta)^2\right)\mathrm{d}\theta$$

$$\text{and}\quad B(\delta) = -\int_{\Theta_\varepsilon(X)^c} \langle\tilde{\nabla}\mathscr{E}_\theta(X), \delta\rangle_{\mathbb{R}^{N\times d}}\mathrm{d}\theta$$

Note that in the expression of $B(\delta)$ the permutation σ_θ involved in $\tilde{\nabla}\mathscr{E}_\theta(X)$ is not necessary unique, and can be chosen arbitrarily.

One has,

$$|\langle\tilde{\nabla}\mathscr{E}_\theta(X), \delta\rangle_{\mathbb{R}^{N\times d}}| \leqslant \|X - Y\circ\sigma^\theta\|_{\mathbb{R}^{N\times d}}\|\delta\|_{\mathbb{R}^{N\times d}}$$

which implies, using Lemma 3

$$|B(\delta)| \leqslant O(\mathrm{Vol}(\Theta_\varepsilon(X)^c)\|\delta\|_{\mathbb{R}^{N\times d}}) = O(\|\delta\|^2_{\mathbb{R}^{N\times d}}) = o(\|\delta\|_{\mathbb{R}^{N\times d}}). \tag{56}$$

Since $(\theta, X) \mapsto \mathscr{E}_\theta(X)$ is continuous and defined on a compact set, it is uniformly continuous, and thus

$$|\mathscr{W}(X_\theta+\delta_\theta, Y_\theta)^2 - \mathscr{W}(X_\theta, Y_\theta)^2| \leqslant C(\delta)$$

where $C(\delta) \to 0$ where $\delta \to 0$. This shows that

$$|A(\delta)| \leqslant \mathrm{Vol}(\Theta_\varepsilon(X)^c)C(\delta) = o(\|\delta\|_{\mathbb{R}^{N\times d}}). \tag{57}$$

Putting together (56) and (57) leads to

$$|\mathscr{E}(X+\delta) - \mathscr{E}(X) - \langle\tilde{\nabla}\mathscr{E}(X), \delta\rangle| = o(\|\delta\|_{\mathbb{R}^{N\times d}})$$

which shows that $\mathscr{E}$ is differentiable with $\nabla\mathscr{E} = \tilde{\nabla}\mathscr{E}$.

Proof of Lipschitzianity of the gradient. For all $\theta \in \Theta_0(X)$, $\nabla\mathscr{E}_\theta(X)$ is continuous and uniformly bounded, and thus $\nabla\mathscr{E}$ is continuous. One has, for $\delta \in \mathbb{R}^{N\times d}$, and denoting $\varepsilon = \|\delta\|$,

$$\nabla\mathscr{E}(X+\delta) - \nabla\mathscr{E}(X) = M(\Theta_\varepsilon(X)) + M(\Theta_\varepsilon(X)^c)$$

$$\text{where}\quad M(U) = \int_U (\nabla\mathscr{E}_\theta(X+\delta) - \nabla\mathscr{E}_\theta(X))\mathrm{d}\theta.$$

One has

$$M(\Theta_\varepsilon(X)) = \int_{\Theta_\varepsilon(X)} \delta_\theta\,\theta\,\mathrm{d}\theta$$

whereas

$$M(\Theta_\varepsilon(X)^c) = \int_{\Theta_\varepsilon(X)^c} \delta_\theta\,\theta\,\mathrm{d}\theta + \int_{\Theta_\varepsilon(X)^c} (Y\circ\tilde{\sigma}_\theta - Y\circ\sigma_\theta)\theta\,\mathrm{d}\theta$$

where $\tilde{\sigma}_\theta = \sigma_{Y_\theta}\circ\sigma^{-1}_{X_\theta+\delta_\theta}$. Using Lemma (3), one has for some constant $C > 0$, $\mathrm{Vol}(\Theta_\varepsilon(X)^c) \leqslant C\|\delta\|_{\mathbb{R}^{N\times d}}$ and hence

$$\|\nabla\mathscr{E}(X+\delta) - \nabla\mathscr{E}(X)\|_{\mathbb{R}^{N\times d}} \leqslant (1 + 2C\|Y\|_{\mathbb{R}^{N\times d}})\|\delta\|_{\mathbb{R}^{N\times d}}$$

which shows that $\nabla\mathscr{E}$ is $(1+2C\|Y\|_{\mathbb{R}^{N\times d}})$-Lipschitz continuous.

References

1. Agueh, M., Carlier, G.: Barycenters in the wasserstein space. SIAM Journal on Mathematical Analysis **43**(2), 904–924 (2011)
2. Averbuch, A., Coifman, R., Donoho, D., Israeli, M., Shkolnisky, Y., Sedelnikov, I.: A framework for discrete integral transformations: II. The 2D discrete Radon transform. SIAM J. Sci. Comput. **30**(2), 785–803 (2008)
3. Benamou, J.D., Brenier, Y.: A computational fluid mechanics solution of the monge-kantorovich mass transfer problem. Numerische Mathematik **84**(3), 375–393 (2000)
4. Benamou, J.D., Froese, B.D., M., O.A.: Numerical solution of the optimal transportation problem via viscosity solutions for the Monge-Ampere equation. CoRR **abs/1208.4873** (2012)
5. Bertsekas, D.: The auction algorithm: A distributed relaxation method for the assignment problem. Annals of Operations Research **14**, 105–123 (1988)
6. Bigot, J., T., K.: Consistent estimation of a population barycenter in the wasserstein space. Preprint arXiv:1212.2562 (2012)
7. Boman, J., Lindskog, F.: Support theorems for the Radon transform and Cramèr-Wold theorems. Journal of Theoretical Probability **22**(3), 683–710 (2009)
8. Bonneel, N., van de Panne, M., Paris, S., Heidrich, W.: Displacement interpolation using lagrangian mass transport. ACM Transactions on Graphics (SIGGRAPH ASIA'11) **30**(6) (2011)
9. Brady, M.L.: A fast discrete approximation algorithm for the radon transform. Journal of Computing **27**(1), 107–119 (1998)
10. Cuturi, M., Doucet, A.: Fast computation of wasserstein barycenters. Preprint arXiv:1212.2562 (2013)
11. Dellacherie, C., Meyer, P.A.: Probabilities and Potential. Math. Stud. 29, North Holland, Amsterdam (1978)
12. Delon, J.: Movie and video scale-time equalization application to flicker reduction. IEEE Transactions on Image Processing **15**(1), 241–248 (2006)
13. Desolneux, A., Moisan, L., Ronsin, S.: A compact representation of random phase and Gaussian textures. In: Proc. the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1381–1384 (2012)
14. Digne, J., Cohen-Steiner, D., Alliez, P., Goes, F., Desbrun, M.: Feature-preserving surface reconstruction and simplification from defect-laden point sets. Journal of Mathematical Imaging and Vision pp. 1–14 (2013)
15. Ferradans, S., Xia, G.S., Peyré, G., Aujol, J.F.: Optimal transport mixing of gaussian texture models. In: Proc. SSVM'13 (2013)
16. Galerne, B., Gousseau, Y., Morel, J.M.: Random phase textures: Theory and synthesis. IEEE Trans. on Image Processing **20**(1), 257–267 (2011)
17. Galerne, B., Lagae, A., Lefebvre, S., Drettakis, G.: Gabor noise by example. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2012) **31**(4), 73:1–73:9 (2012)
18. H. W. Kuhn: The Hungarian method of solving the assignment problem. Naval Res. Logistics Quart. **2**, 83–97 (1955)
19. Haker, S., Zhu, L., Tannenbaum, A., Angenent, S.: Optimal mass transport for registration and warping. International Journal of Computer Vision **60**(3), 225–240 (2004)
20. Helgason, S.: The Radon Transform. Birkhauser, Boston (1980)
21. Kantorovich, L.: On the transfer of masses (in russian). Doklady Akademii Nauk **37**(2), 227–229 (1942)
22. Matusik, W., Zwicker, M., Durand, F.: Texture design using a simplicial complex of morphable textures. ACM Transactions on Graphics **24**(3), 787–794 (2005)
23. McCann, R.J.: A convexity principle for interacting gases. advances in mathematics **128**(1), 153–179 (1997)
24. Mérigot, Q.: A multiscale approach to optimal transport. Computer Graphics Forum **30**(5), 1583–1592 (2011)

25. Papadakis, N., Peyré, G., Oudet, E.: Optimal transport with proximal splitting. SIAM Journal on Imaging Sciences **7**(1), 212?–238 (2014)
26. Pitié, F., Kokaram, A.C., Dahyot, R.: N-dimensional probability density function transfer and its application to color transfer. In: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, vol. 2, pp. 1434–1439. IEEE (2005)
27. Rabin, J., Delon, J., Gousseau, Y.: Removing artefacts from color and contrast modifications. IEEE Transactions on Image Processing **20**(11), 3073–3085 (2011)
28. Rabin, J., Peyré, G., Delon, J., Bernot, M.: Wasserstein barycenter and its application to texture mixing. In: Scale Space and Variational Methods in Computer Vision (SSVM'11), vol. 6667, pp. 435–446 (2011)
29. Reinhard, E., Pouli, T.: Colour spaces for colour transfer. In: Proceedings of the Third international conference on Computational color imaging, CCIW'11, pp. 1–15. Springer-Verlag, Berlin, Heidelberg (2011)
30. Rubner, Y., Tomasi, C., Guibas, L.: A metric for distributions with applications to image databases. In: IEEE International Conference on Computer Vision (ICCV'98), pp. 59–66 (1998)
31. Solodov, M.: Incremental gradient algorithms with stepsizes bounded away from zero. Computational Optimization and Applications **11**(1), 23–35 (1998)
32. Villani, C.: Topics in Optimal Transportation. Graduate Studies in Mathematics Series. American Mathematical Society (2003)