Optimal Transport for Manifold-Valued Images

Jan Henrik Fitschen[1], Friederike Laus[1], and Bernhard Schmitzer[2]

[1] Department of Mathematics, University of Kaiserslautern, Germany
{fitschen,friederike.laus}@mathematik.uni-kl.de
[2] Institute for Computational and Applied Mathematics University of Münster, Germany
schmitzer@uni-muenster.de

Abstract. We introduce optimal transport-type distances for manifold-valued images. To do so we lift the initial data to measures on the product space of image domain and signal space, where they can then be compared by optimal transport with a transport cost that combines spatial distance and signal discrepancy. Applying recently introduced 'unbalanced' optimal transport models leads to more natural results. We illustrate the benefit of the lifting with numerical examples for interpolation of color images and classification of handwritten digits.

1 Introduction

The optimal transport (OT) problem has found various applications in signal and image processing, computer vision and machine learning [13,7,8,15]. In imaging applications one typically faces two problems: Standard OT can only compare measures of equal mass which is often an unnatural assumption. Further, the involved discretized transport problems are very high-dimensional and challenging to solve. The former problem has for instance been addressed by the Hellinger-Kantorovich (HK) distance (or Wasserstein-Fisher-Rao distance) [10,5,12], which allows to compare 'unbalanced' measures of different mass. This metric does not only extend to measures of different total mass, but yields often also more reasonable results on balanced measures by removing artifacts where small portions of mass would otherwise have to be moved very far. For numerically solving OT problems a broad family of methods has been devised, such as the network simplex [1] or a fluid-dynamic formulation [3]. Another choice is the entropy regularization technique and the Sinkhorn scaling algorithm [7], which has been extended to 'unbalanced' problems in [6].

The OT distance – and the HK distance – induce a metric on non-negative, scalar signals. Recently, the question how to define meaningful transport-type metrics on multi-channel data has arisen. For instance, in [8] OT has been extended to RGB color images, and in [15] so-called TL^p-distances are introduced, a transport-type metric over vector valued signals. It is illustrated that these distances remain sensitive to high-frequency signal oscillations, but are at the same time robust to deformations such as translations, thus uniting the advantages of L^p and Wasserstein distances. In both approaches the original non-scalar data is transformed to a non-negative scalar measure on the product space of image domain and signal space, where it can be compared in a classical OT framework.

Contribution and Outline. In this article we build on the work of [8] and present a framework for transport-type metric on multi-channel and manifold valued data. Also in our approach, signals will be lifted to measures on the product space of image domain and signal space, and then compared with transport-type metrics, using a cost function combining spatial distance and signal discrepancy. For comparison we use the HK distance, which can compare non-normalized images and yields more natural assignments. To solve the resulting, high-dimensional optimization problems we employ efficient diagonal scaling algorithms based on entropy regularization [6] that generalize the well-known Sinkhorn algorithm.

In our numerical examples, we apply our framework to color images in different color spaces by defining three lifting models with different interpretations of the mass and signal component and show that they are adapted for different types of images. The geodesic structure of the OT and HK distances can be used to compute image interpolations. For this we propose a back projection map that takes an intermediate lifted measures back to an intermediate manifold valued image. To showcase the potential for machine learning applications, we further apply the idea to classification of handwritten digits of the MNIST database. The sample data is initially scalar and no lifting would be required. We demonstrate that lifting, based on features extracted from the scalar data, yields improved performance.

The article is organized as follows: In Sect. 2 we introduce the generic mathematical framework of our model. Sect. 3 discusses discretization and optimization. Application of our model to color images and the MNIST handwritten digits database with corresponding numerical results are presented in Sect. 4.

Relation to [8] and [15]. The framework presented in this article is more general compared than the approach of [8] and our optimization scheme extends to more complex signal spaces. Relative to [15] we use unbalanced transport for comparison and introduce the back projection map. Moreover, [15] mostly uses a fixed reference measure with constant density for lifting. We put more emphasis on this choice, in particular we extract the measure from the signal, which is an important part of our model. Conversely, for the MNIST example we propose that the signal can be extracted from the measure.

2 Unbalanced Transport for Manifold-Valued Signals

2.1 Wasserstein and Hellinger-Kantorovich Distances

Let $X \subset \mathbb{R}^d$ be the image domain, e.g. $[0,1]^2$, and $\mathcal{M}$ the signal space, e.g. an appropriate color space. Further, let $d_X(x_0, x_1) = \|x_0 - x_1\|_2$ be the Euclidean distance on X and let $d_\mathcal{M}$ be a suitable metric on $\mathcal{M}$. We denote by $\mathcal{Y} = X \times \mathcal{M}$ the product space which we endow with the metric $d_\mathcal{Y}^2((x_0, m_0), (x_1, m_1)) = d_X^2(x_0, x_1) + \lambda^2 \cdot d_\mathcal{M}^2(m_0, m_1)$, where $\lambda \geq 0$ is a relative weighting parameter. We assume that X, $\mathcal{M}$ and $\mathcal{Y}$ are compact, complete and equipped with their Borel σ-algebras. For a measurable space Z (for example X, $\mathcal{Y}$ or $\mathcal{Y} \times \mathcal{Y}$), we denote by $\mathcal{P}(Z)$ the set of non-negative Radon measures over Z and by $\mathcal{P}_1(Z) \subset \mathcal{P}(Z)$ the set of probability measures. The Dirac delta at $z \in Z$ is denoted by δ_z. For two measurable spaces Z_1, Z_2, a measure $\mu \in \mathcal{P}(Z_1)$ and a measurable map $f\colon Z_1 \to Z_2$, the *push-forward* (or image measure) $f_\# \mu$ of μ under f is given by $(f_\# \mu)(\sigma) = \mu(f^{-1}(\sigma))$ for all measurable sets $\sigma \subset Z_2$.

Let $\mathrm{pr}_{Z,0}\colon Z^2 \to Z$, $(z_0, z_1) \mapsto z_0$ and similarly $\mathrm{pr}_{Z,1}(z_0, z_1) = z_1$ be projections from Z^2 to the first and second component. For a measure $\pi \in \mathcal{P}(Z^2)$ the first and second marginal are then given by $\mathrm{pr}_{Z,i\sharp}\, \pi$, $i \in \{0, 1\}$ respectively.

Let $\mu_i \in \mathcal{P}_1(X)$, $i \in \{0, 1\}$. The set

$$\Pi_X(\mu_0, \mu_1) = \left\{\pi \in \mathcal{P}(X^2)\colon\ \mathrm{pr}_{X,i\sharp}\, \pi = \mu_i \text{ for } i \in \{0, 1\}\right\} \tag{2.1}$$

is called the set of *transport plans* between μ_0 and μ_1. Every $\pi \in \Pi_X(\mu_0, \mu_1)$ describes a rearrangement of the mass of μ_0 into μ_1. The *2-Wasserstein distance* over X between μ_0 and μ_1 is given by

$$d^2_{W,X}(\mu_0, \mu_1) = \inf_{\pi \in \Pi_X(\mu_0,\mu_1)} \int_{X^2} d^2_X(x_0, x_1)\, \mathrm{d}\pi(x_0, x_1). \tag{2.2}$$

This means we are looking for the cheapest plan π, where the cost of moving one unit of mass from x_0 to x_1 is given by $d^2_X(x_0, x_1)$. The Wasserstein distance $d_{W,X}$ is a metric over $\mathcal{P}_1(X)$, for a thorough introduction to this topic we refer to [16].

The distance $d_{W,X}$ only allows comparison between normalized measures: otherwise $\Pi_X(\mu_0, \mu_1)$ is empty and $d_{W,X}(\mu_0, \mu_1) = +\infty$. This is remedied by the Hellinger-Kantorovich distance, as introduced e.g. in [12], which allows creation and annihilation of mass and hence induces a (finite) metric over all non-negative measures $\mathcal{P}(X)$. For a parameter $\kappa > 0$ the Hellinger-Kantorovich distance is given by [12, Sects. 6–7]

$$d^2_{HK,X}(\mu_0, \mu_1) = \kappa^2 \inf_{\pi \in \mathcal{P}(X^2)} \sum_{i=0}^{1} \mathrm{KL}(\mathrm{pr}_{X,i\sharp}\, \pi | \mu_i) + \int_{X^2} c_{X,\kappa}(x_0, x_1)\, \mathrm{d}\pi(x_0, x_1) \tag{2.3}$$

where

$$c_{X,\kappa}(x_0, x_1) = \begin{cases} -\log\left([\cos(d_X(x_0, x_1)/\kappa)]^2\right) & \text{if } d_X(x_0, x_1) < \kappa \cdot \frac{\pi}{2} \\ +\infty & \text{else,} \end{cases} \tag{2.4}$$

and $\mathrm{KL}_X(\mu|\nu)$ denotes the Kullback-Leibler (KL) divergence from ν to μ. Compared to (2.2) the constraint $\mathrm{pr}_{X,i\sharp}\, \pi = \mu_i$, $i \in \{0, 1\}$ is relaxed. The plan π can now be any non-negative measure and the discrepancy between $\mathrm{pr}_{X,i\sharp}\, \pi$ and μ_i is penalized by the KL divergence. This implies in particular that $d_{HK,X}(\mu_0, \mu_1)$ is finite when $\mu_0(X) \neq \mu_1(X)$. Additionally, the cost d^2_X is replaced by $c_{X,\kappa}$. Note that $c_{X,\kappa}(x_0, x_1) = +\infty$ if $d_X(x_0, x_1) \geq \kappa \cdot \pi/2$. Thus, beyond this distance no transport occurs and mass growth completely takes over. The precise form of $c_{X,\kappa}$ is determined by an equivalent fluid-dynamic-type formulation [12]. The parameter κ balances between transport and mass changes. As $\kappa \to 0$, $d_{HK,X}/\kappa$ converges towards the Hellinger distance, which is purely local and does not involve transport. For $\kappa \to \infty$ one finds $d_{HK,X}(\mu_0, \mu_1) \to d_{W,X}(\mu_0, \mu_1)$ [12, Sect. 7.7].

When (X, d_X) is a length space, this holds for $(\mathcal{P}_1(X), d_{W,X})$ as well. Let μ_0, $\mu_1 \in \mathcal{P}_1(X)$ and let π be a corresponding minimizer of (2.2). A geodesic between μ_0 and μ_1 can be reconstructed from π. Intuitively, a mass particle at $\pi(x_0, x_1)$ has to travel with constant speed along the geodesic from x_0 to x_1 (which is a straight line in $\mathbb{R}^d$). This can be formalized as follows: for $t \in [0, 1]$ let

$$\gamma_t \colon X \times X \to X, \qquad (x_0, x_1) \mapsto (1-t) \cdot x_0 + t \cdot x_1\,. \tag{2.5}$$

That is, $t \mapsto \gamma_t(x_0, x_1)$ describes the geodesic between x_0 and x_1 in X. Then a geodesic between μ_0 and μ_1 is given by

$$[0, 1] \ni t \mapsto \mu_t = \gamma_{t\#}\pi\,. \tag{2.6}$$

This is the *displacement interpolation* [16] between μ_0 and μ_1. It often provides a more natural interpolation than the naive linear trajectory $[0, 1] \ni t \mapsto (1 - t) \cdot \mu_0 + t \cdot \mu_1$. The famous Benamou-Brenier formula [3] is an equivalent fluid-dynamic-type reformulation of (2.2) directly in terms of finding the displacement interpolation.

Similarly, $(\mathcal{P}(X), d_{HK,X})$ is a length space and geodesics can be constructed from optimal couplings π in (2.3). The construction is more involved, since particles change their mass and speed while travelling. We refer to [12, Part II] for a detailed elaboration of $d_{HK,X}$ geodesics. There also is a fluid-dynamic-type formulation for $d_{HK,X}$ [10,5,12], in which mass changes are penalized by the Hellinger (or Fisher-Rao) distance. This equivalence determines the precise form of (2.4). An illustration of displacement interpolations is given in Fig. 1.

2.2 Manifold-Valued Images, Lifted Measures and Distances

Next, we extend the transport-type metrics to $\mathcal{M}$-valued signals. First, observe that in complete analogy to Sect. 2.1, $\Pi_{\mathcal{Y}}$, $\mathrm{KL}_{\mathcal{Y}}$ and $c_{HK,\mathcal{Y}}$ can be defined over the metric space $(\mathcal{Y}, d_{\mathcal{Y}})$. Consequently, the Wasserstein distance $d_{W,\mathcal{Y}}$ and Hellinger-Kantorovich distance $d_{HK,\mathcal{Y}}$ can be constructed over $\mathcal{P}_1(\mathcal{Y})$ and $\mathcal{P}(\mathcal{Y})$. For example, for ν_0, $\nu_1 \in \mathcal{P}_1(\mathcal{Y})$ we find

$$d^2_{W,\mathcal{Y}}(\nu_0, \nu_1) = \inf_{\pi \in \Pi_{\mathcal{Y}}(\nu_0,\nu_1)} \int_{\mathcal{Y}^2} d^2_{\mathcal{Y}}\big((x_0, m_0), (x_1, m_1)\big)\, \mathrm{d}\pi\big((x_0, m_0), (x_1, m_1)\big). \tag{2.7}$$

Now, the key is to lift the original $\mathcal{M}$-valued signals over X to non-negative measures on the product space $\mathcal{Y} = X \times \mathcal{M}$, where they can then be compared with $d_{W,\mathcal{Y}}$ and $d_{HK,\mathcal{Y}}$. Let $f_i \colon X \to \mathcal{M}$, $i \in \{0, 1\}$ be two (measurable) $\mathcal{M}$-valued images that we want to compare and let $\mu_i \in \mathcal{P}(X)$ be two corresponding

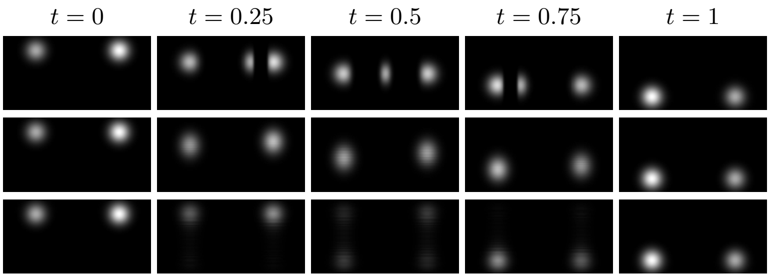

Fig. 1. Geodesics in $(\mathcal{P}_1(X), d_{W,X})$ and $(\mathcal{P}(X), d_{HK,X})$. *Top row:* Wasserstein geodesic between two pairs of Gaussians with different mass. To compensate the local difference, mass is sent from the right to the left. *Middle row:* Hellinger-Kantorovich geodesic for $\kappa = 60$. The mass difference between upper and lower Gaussians is compensated locally by creating and shrinking mass, leading to a more natural interpolation. Note that the left and right Gaussian are travelling at slightly different speeds and are slightly ellipsoidal during transport, which is characteristic for d_{HK} geodesics. *Bottom row:* d_{HK} geodesic for $\kappa = 16$. A lot of the differences between the two images is now compensated purely by growth and shrinkage of mass.

measures. The choice of the measures μ_i is an important part of the model and we will describe this choice in detail for each model in Sect. 4.1.

From the pairs (f_i, μ_i), $i \in \{0, 1\}$, we generate two measures on $\mathcal{Y}$ as follows:

$$\nu_i = F_{i\#}\mu_i, \quad \text{where } F_i\colon X \to \mathcal{Y}, \quad x \mapsto \big(x, f_i(x)\big) \tag{2.8}$$

For example, if $\mu_0 = \delta_x$ then $\nu_0 = \delta_{(x,f_0(x))}$. That is, a Dirac measure at $x \in X$ is lifted to a Dirac at $(x, f_0(x)) \in \mathcal{Y}$. The signal f_i becomes encoded in the position of the mass of ν_i (i.e. we only 'care' about the values of f_i μ_i-a.e.).

Let $F\colon (x_0, x_1) \mapsto \big(F_0(x_0), F_1(x_1)\big) = \big((x_0, f_0(x_0)), (x_1, f_1(x_1))\big)$. It is then a simple exercise to show that $F_\# \Pi_X(\mu_0, \mu_1) = \Pi_{\mathcal{Y}}(F_{0\#}\mu_0, F_{1\#}\mu_1)$. Consequently, for lifted measures $\nu_i = F_{i\#}\mu_i$ the lifted Wasserstein distance can be written as

$$d^2_{W,\mathcal{Y}}(\nu_0, \nu_1) = \inf_{\pi \in \Pi_X(\mu_0,\mu_1)} \int_{X^2} \big(d^2_X(x_0, x_1) + \lambda^2 d^2_{\mathcal{M}}\big(f_0(x_0), f_1(x_1)\big)\big)\, \mathrm{d}\pi(x_0, x_1). \tag{2.9}$$

This implies that the lifted distance $d_{W,\mathcal{Y}}$ between lifted signals can be computed by a transport problem over X, where the transport cost between x_0 and x_1 not only depends on the spatial distance $d_X(x_0, x_1)$, but also on the 'signal distance' $d_{\mathcal{M}}(f_0(x_0), f_1(x_1))$. An analogous interpretation holds for the lifted Hellinger-Kantorovich distance. The authors of [15] provide some intuition for lifted distances $(\mathcal{Y}, d_{W,\mathcal{Y}})$ and show that (2.9) defines a distance over (signal,measure)-pairs (f, μ). We will illustrate these lifted distances and demonstrate their benefit for meaningful image registration and enhanced classification scores for various example models throughout Sect. 4.

Again, in analogy to Sect. 2.1, when $(\mathcal{Y}, d_{\mathcal{Y}})$ is a length space, so are $(\mathcal{P}_1(\mathcal{Y}), d_{W,\mathcal{Y}})$ and $(\mathcal{P}(\mathcal{Y}), d_{HK,\mathcal{Y}})$. Accordingly, for two marginals ν_0, $\nu_1 \in \mathcal{P}_1(\mathcal{Y})$ (or

$\mathcal{P}(\mathcal{Y})$) one can construct geodesics $t \mapsto \nu_t$ similar to (2.6) (or for d_{HK}). The main difference is that the map γ_t, describing geodesics in X, (2.5) has to be replaced by geodesics on $(\mathcal{Y}, d_\mathcal{Y})$. In the lifted geodesics, mass is travelling both in spatial direction X as well as the 'signal' direction $\mathcal{M}$.

We want to use such geodesics to interpolate between pairs of signals (f_i, μ_i) that we lift to measures ν_i via (2.8), $i \in \{0, 1\}$. However, an intermediate point ν_t, $t \in (0, 1)$ on the geodesic between ν_0 and ν_1, cannot always be written as a lifting of an intermediate pair (f_t, μ_t). Intuitively, this is because at time t several mass particles might occupy the same spatial location $x \in X$, but at different signal positions m, $m' \in \mathcal{M}$ and such a constellation cannot be obtained by a push-forward from $\mathcal{P}(X)$ as in (2.8). We propose to resolve such overlaps by picking for each location in $x \in X$ the barycenter $\overline{m} \in \mathcal{M}$ w.r.t. $d_\mathcal{M}$ of all signal values m and m' that can be found at this location. Let $\rho \in \mathcal{P}_1(\mathcal{M})$ describe a signal distribution of lifted mass particles in $\mathcal{Y}$ 'over' a given location $x \in X$. The barycenter of ρ is defined as

$$\mathrm{Bar}(\rho) = \underset{\overline{m} \in \mathcal{M}}{\mathrm{argmin}} \int_\mathcal{M} d^2_\mathcal{M}(\overline{m}, m)\, \mathrm{d}\rho(m)\,. \tag{2.10}$$

In this article we assume that a unique minimizer exists. When $\mathcal{M}$ is a convex subset of $\mathbb{R}^n$ the barycenter is given by the center of mass. To construct (f, μ) from a given $\nu \in \mathcal{P}(\mathcal{Y})$ we first set

$$\mu = \mathsf{P}_\# \nu \qquad \text{where} \qquad \mathsf{P}\colon \mathcal{Y} \to X, \quad (x, m) \mapsto x\,. \tag{2.11}$$

That is, at every point $x \in X$, μ gathers all the mass of ν in the fibre $\{x\} \times \mathcal{M}$. Then, by the disintegration theorem [2, Thm. 5.3.1], there is a family of probability measures ρ_x for all $x \in X$ (unique μ-a.e.) such that we can write

$$\int_\mathcal{Y} \phi(x, m)\, \mathrm{d}\nu(x, m) = \int_X \left(\int_\mathcal{M} \phi(x, m)\, \mathrm{d}\rho_x(m) \right) \mathrm{d}\mu(x) \tag{2.12}$$

for any measurable $\phi\colon \mathcal{Y} \to [0, +\infty]$. ρ_x can be thought of as describing how the mass of ν in the fibre $\{x\} \times \mathcal{M}$ is distributed. Now, we set $f(x) = \mathrm{Bar}(\rho_x)$, which is well-defined μ-almost everywhere. This signal f is the best point-wise approximation of the lifted measure ν in the sense of (2.10). We call (f, μ) the *back projection* of ν. Note that if ν is in fact a lifting of some (f, μ), then $\rho_x = \delta_{f(x)}$ μ-a.e. and (f, μ) are recovered (μ-a.e.) by back projection.

3 Discretization and Optimization

For our numerical experiments we assume that all measures μ_i are concentrated on a discrete Cartesian pixel grid $\mathbf{X} = \{x_1, \ldots, x_N\} \subset X \subset \mathbb{R}^2$ and we only care about the values of signals f_i on $\mathbf{X}$. Thus, all feasible π in (2.9) are concentrated on $\mathbf{X}^2$ and (2.9) becomes a finite dimensional problem. It can be written as

$$d^2_{W,\mathcal{Y}}(F_{0\#}\mu_0, F_{1\#}\mu_1) = \inf_{\boldsymbol{\pi} \in \boldsymbol{\Pi}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1)} \langle \boldsymbol{d}, \boldsymbol{\pi} \rangle\,, \text{ where } \langle \boldsymbol{d}, \boldsymbol{\pi} \rangle \overset{\text{def.}}{=} \sum_{j,k=1}^N \boldsymbol{d}_{j,k}\, \boldsymbol{\pi}_{j,k}\,, \tag{3.1}$$

with discrete vectors $\boldsymbol{\mu}_i \in \mathbb{R}^N$, $(\boldsymbol{\mu}_i)_j = \mu_i(\{x_j\})$, discrete couplings $\boldsymbol{\Pi}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1) = \{\boldsymbol{\pi} \in \mathbb{R}_+^{N\times N} \colon \boldsymbol{\pi}\,\mathbf{1} = \boldsymbol{\mu}_0, \boldsymbol{\pi}^\top\mathbf{1} = \boldsymbol{\mu}_1\}$ and $\boldsymbol{d}_{j,k} = d_X^2(x_j, x_k) + \lambda^2 d_{\mathcal{M}}^2\big(f_0(x_j), f_1(x_k)\big)$. Here, $\mathbf{1} \in \mathbb{R}^N$ is the vector with all entries being 1, $\boldsymbol{\pi}\,\mathbf{1}$, $\boldsymbol{\pi}^\top\mathbf{1}$ give the column and row sums of $\boldsymbol{\pi}$, which is the discrete equivalent of $\mathrm{pr}_{X,i\sharp}\,\pi$. Similarly, since $\mathrm{KL}(\mu|\nu) = +\infty$ if $\mu \not\ll \nu$, all feasible π in the unbalanced problem (2.3) are concentrated on $\mathbf{X}^2$. The discrete unbalanced equivalent of (2.9) becomes

$$d_{HK,\mathcal{Y}}^2(F_{0\#}\mu_0, F_{1\#}\mu_1) = \kappa^2 \inf_{\boldsymbol{\pi}\in\mathbb{R}_+^{N\times N}} \mathbf{KL}(\boldsymbol{\pi}\,\mathbf{1}|\boldsymbol{\mu}_0) + \mathbf{KL}(\boldsymbol{\pi}^\top\mathbf{1}|\boldsymbol{\mu}_1) + \langle \boldsymbol{c}, \boldsymbol{\pi}\rangle\,, \tag{3.2}$$

where $\boldsymbol{c}_{j,k} = c_{HK,\mathcal{Y}}\big((x_j, f_0(x_j)), (x_k, f_1(x_k))\big)$, analogous to $\boldsymbol{d}$, and $\mathbf{KL}$ is the discrete Kullback-Leibler divergence. Note that this approach does not require discretization of the signal space $\mathcal{M}$. Once an optimal $\boldsymbol{\pi}$ for the finite dimensional problems (3.1) or (3.2) is obtained, it can be used to construct the geodesic between $F_{0\#}\mu_0$ and $F_{1\#}\mu_1$ with the lifted variants of (2.6) (or for d_{HK}), see above. These geodesics consist of a finite number of Dirac measures travelling smoothly through $\mathcal{Y}$.

Now, we describe the discretized back projection. To generate an intermediate image (f_t, μ_t), living on the discrete grid $\mathbf{X}$, we proceed as follows: the mass of any travelling Dirac at location $(x, m) \in \mathcal{Y}$ is distributed to the closest four pixels in $\mathbf{X}$ according to bilinear interpolation. In this way, we obtain for each pixel $x_j \in \mathbf{X}$ a total mass, corresponding to $\mu_t(\{x_j\}) = (\boldsymbol{\mu}_t)_j$, (2.11), and a distribution over $\mathcal{M}$, corresponding to ρ_x, (2.12). The barycenter of this distribution, (2.10), yields the discrete backprojected signal $f_t(x_j)$.

To solve problems (3.1) and (3.2) we employ the entropy regularization approach for optimal transport [7] and unbalanced transport [6]. For a (small) positive parameter $\varepsilon > 0$ we regularize the problems by adding the term $\varepsilon \cdot \mathbf{KL}(\boldsymbol{\pi}|\boldsymbol{I})$, where $\boldsymbol{I} \in \mathbb{R}^{N\times N}$ is the matrix with all entries being 1 and by a slight abuse of notation $\mathbf{KL}$ denotes the discrete KL-divergence extended to matrices. The regularized variant of (3.1) can be solved with the Sinkhorn algorithm [7], the regularized version of (3.2) with a slightly more general (but equally simple) scaling algorithm [6]. In our examples, N is of the order of 10^4, hence working on the full grid $\mathbf{X} \times \mathbf{X}$ is infeasible. To obtain good approximations to the original problems (3.1) and (3.2) we want to choose a small ε, which leads to several numerical issues. To remedy these problems we employ the numerical scheme described in [14]. In particular this allows setting ε small enough to make the induced entropic smoothing practically negligible and uses sparse approximations of $\mathbf{X}^2$ to reduce the required memory and runtime.

In [8] a Benamou-Brenier-type formula [3] was used to solve problem (2.7) for RGB-images (cf. Sect. 4.1) with a particular formulation that required only three points to discretize $\mathcal{M}$ (corresponding to three color channels). However, it would be challenging to generalize this approach to other, higher-dimensional $\mathcal{M}$, since it would entail discretizing the high-dimensional space $\mathcal{Y} = X \times \mathcal{M}$.

4 Examples and Numerical Results

4.1 Color Images

Let $\mathbf{X} \subset \mathbb{R}^2$ be the discrete image domain (cf. Sect. 3) and let $g\colon \mathbf{X} \to \mathcal{C}$ be a color image, where $\mathcal{C}$ is either the RGB or HSV color space (for details on the color spaces we refer to [9]). In the following, we present three different ways how to choose $\mathcal{M}$ and how to generate the pair (f, μ) from g. In the first two models μ takes the form $\mu = \sum_{x \in \mathbf{X}} w(x) \cdot \delta_x$, so μ can be specified by fixing the weighting function $w\colon \mathbf{X} \to \mathbb{R}_+$.

rgb-cube: Let $\mathcal{C} = [0,1]^3$ be the RGB color space. In this model we choose $\mathcal{M} = \mathcal{C}$ and $d_{\mathcal{M}}$ is the Euclidean distance on $\mathcal{M}$. We set $f = g$ and $w(x) = 1$ for all $x \in \mathbf{X}$, i.e. μ is the 'uniform' counting measure over the pixels. So every pixel gets lifted to a point determined by its RGB values with mass 1.

hsv-disk: Let $\mathcal{C} = S^1 \times [0,1]^2$ represent the HSV color space and let $g = (h, s, v)$ be a triplet of functions specifying hue, saturation and value of each pixel. In this model we choose $\mathcal{M} = S^1 \times [0,1]$ and set $f = (h, s)$ and $w = v$. For (h_0, s_0), $(h_1, s_1) \in \mathcal{M}$ the metric is given by

$$d^2_{\mathcal{M}}((h_0, s_0), (h_1, s_1)) = \left\| \begin{pmatrix} s_0 \cos(h_0) \\ s_0 \sin(h_0) \end{pmatrix} - \begin{pmatrix} s_1 \cos(h_1) \\ s_1 \sin(h_1) \end{pmatrix} \right\|_2^2,$$

which is the Euclidean distance in polar coordinates. In this model, hue and saturation of each pixel are transformed into lifted coordinates and the value channel is transformed into mass. Black pixels are not assigned any mass. This model is suited for scenarios where intensity puts the emphasis on certain details, e.g. bright objects on a dark background.

rgb-triple: Let again $\mathcal{C} = [0,1]^3$ be the RGB color space. This model requires a slight extension of the lifting framework and (2.8), as every pixel is transformed into three Dirac masses. As in rgb-cube we choose $\mathcal{M} = [0,1]^3$. Let $g = (\mathsf{r}, \mathsf{g}, \mathsf{b})$ be a function triplet specifying an RGB image. We define the lifted measure as

$$\nu = \sum_{x \in \mathbf{X}} \mathsf{r}(x) \cdot \delta_{(x,(1,0,0)^\top)} + \mathsf{g}(x) \cdot \delta_{(x,(0,1,0)^\top)} + \mathsf{b}(x) \cdot \delta_{(x,(0,0,1)^\top)}.$$

To reconstruct a color image from a (discrete) measure $\nu \in \mathcal{P}(\mathcal{Y})$ we map a Dirac $\rho \cdot \delta_{(x,(z_1,z_2,z_3))}$ to the color $(\mathsf{r}, \mathsf{g}, \mathsf{b}) = (\rho\, z_1, \rho\, z_2, \rho\, z_3)$. This is a reformulation of the color transport model of [8] in our framework. It is particularly suited in cases where we want to model additive mixing of colors. In Fig. 2 we visualize geodesics in $(\mathcal{M}, d_{\mathcal{M}})$ for the models rgb-cube and hsv-disk in the RGB cube and the HSV cylinder. For example, the trajectory from blue to red in both models goes via pink, as expected. But the precise transition varies. Fig. 3 shows the transport between simple mixtures of Gaussians of different colors to illustrate the behavior of our model. The first two rows show how the weighting between color transport and spatial transport influences the result. The third row depicts the result of the back projection in case of superposition of two Gaussians. Finally, in the last two rows the fundamentally different behavior of the models hsv-disk and

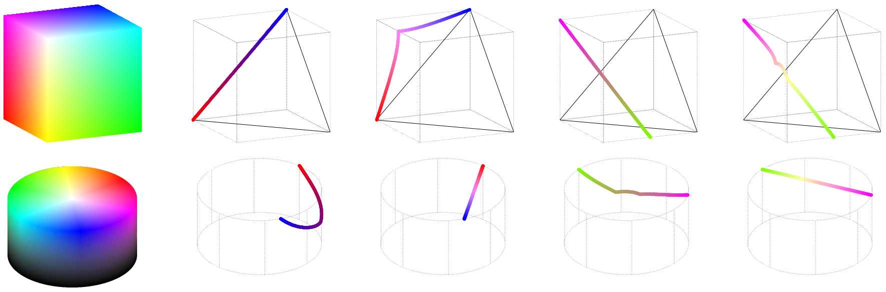

Fig. 2. Geodesics in the models rgb-cube (second and fourth column) and hsv-disk (third and fifth column), visualized in the RGB cube (top) and the HSV cylinder (bottom). Trajectories from 'red' $(\mathsf{r},\mathsf{g},\mathsf{b}) = (1,0,0)$ to 'blue' $(\mathsf{r},\mathsf{g},\mathsf{b}) = (0,0,1)$ (second and third column) and from 'pink' $(\mathsf{r},\mathsf{g},\mathsf{b}) = (1,0,1)$ to 'green' $(\mathsf{r},\mathsf{g},\mathsf{b}) = (\frac{1}{2},1,0)$ (fourth and fifth column).

rgb-triple is visible, here the rgb-triple model leads to a division of color into the single color channels.

Next, we provide three examples for real images in Fig. 4. The examples are chosen in order to illustrate which model is suited best for which kind of images. For standard photographies, the rgb-cube model is suited best. As an example, we show in the top row the interpolation between two photos of a tree, taken in summer and autumn.[3] For the second row we use the same optimal π, but during interpolation mass particles only travel in signal direction, staying at the same spatial location. Thus, only the color is transformed while the geometry of the image remains fixed (cf. [15]). Both cases yield realistic interpolations, despite slight artifacts in the first row. This is expected, as optimal transport does not explicitly enforce spatial regularity of the optimal matching. For geometrically aligned images as in our example, these artifacts can be removed in a local post-processing step. Future work will include studying additional terms to enforce more spatially regular interpolations between images with substantially different geometries. In the hsv-disk model, the mass of a pixel depends on its intensity. As a consequence, the model is well suited for images showing bright objects on a dark background. An example for such kind of images are fireworks, and the third and fourth row Fig. 4 show the results obtained by transporting such images in the hsv-disk model. In both cases, color and shape are nicely interpolated during the transport. In the rgb-triple model, each color channel is treated separately, which allows the mass to travel through the channels. Such a color decomposition naturally occurs in the spectral decomposition of light. The fifth row of Fig. 4 gives the interpolation between an image with pure white light and an image showing the single colors of the spectrum obtained with a prism.[4]

[3] The tree images were kindly provided by Dr. J. Hagelüken.

[4] The images were kindly provided by N. Spiker.

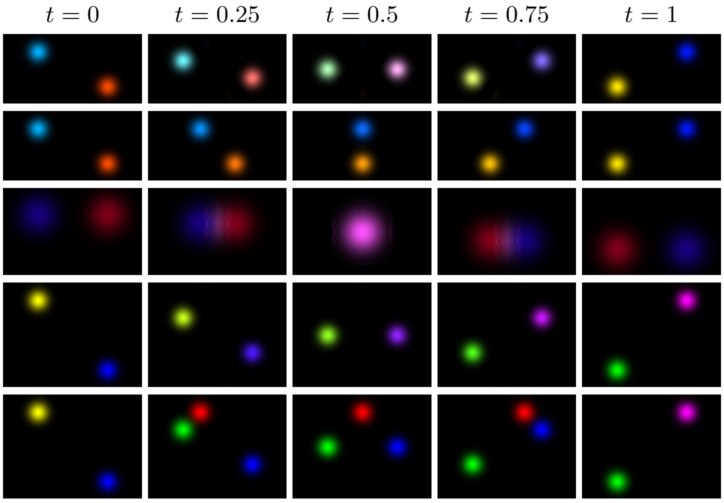

Fig. 3. Back projection of several geodesics between mixtures of Gaussians in $d_{W,\mathcal{Y}}$ for the color model hsv-disk (and rgb-triple, last row). *First row:* $\lambda \approx 0$, transport cost depends essentially only on spatial location, regardless of colors. *Second row:* with $\lambda = 50$ a more natural color assignment is obtained. *Third row:* 'collision' of a blue and a red Gaussian results in a pink Gaussian via back projection. *Fourth and fifth row:* comparison between color models: hsv-disk gradually interpolates the hue, rgb-triple decomposes each pixel into elementary colors and rearranges these components.

4.2 MNIST Handwritten Digits

In this section we present an example for a machine learning application on the MNIST handwritten digits dataset [11]. The dataset consists of 28 × 28 pixel gray level images of handwritten digits $\{0, 1, \ldots, 9\}$, which we interpret as probability measures in $\mathcal{P}_1(X)$. Note that despite using d_{HK} we choose to normalize all images before comparison because we do not want *global* mass differences to influence the distance, but only *local* discrepancies.

Since the original samples already lie in $\mathcal{P}_1(X)$, they can directly be compared with $d_{W,X}$ or $d_{HK,X}$. A priori, there is no need for a signal manifold $\mathcal{M}$ or lifting. Nevertheless, for an image $\mu \in \mathcal{P}_1(X)$, we propose to interpret a locally smoothed Hessian of the image as $\mathbb{R}^{2\times 2} = \mathcal{M}$-valued signal f, metrized by the Frobenius norm. Since the Hessian represents local image curvature, the signal f can be thought of as encoding the local orientation of the lines in μ. We demonstrate that by using this additional information, the lifted distances $d_{W,\mathcal{Y}}$ and $d_{HK,\mathcal{Y}}$ can discriminate more accurately between different digits. Fig. 5 shows the improved performance in nearest neighbour retrieval and a clearer class separation in a multi-dimensional scaling representation. Note that both the lifting, as well as switching from standard to unbalanced transport improve the performance. The approach to generate the signal f from μ can be seen as complimentary to the color models (Sect. 4.1) and illustrates the flexibility of the lifting approach.

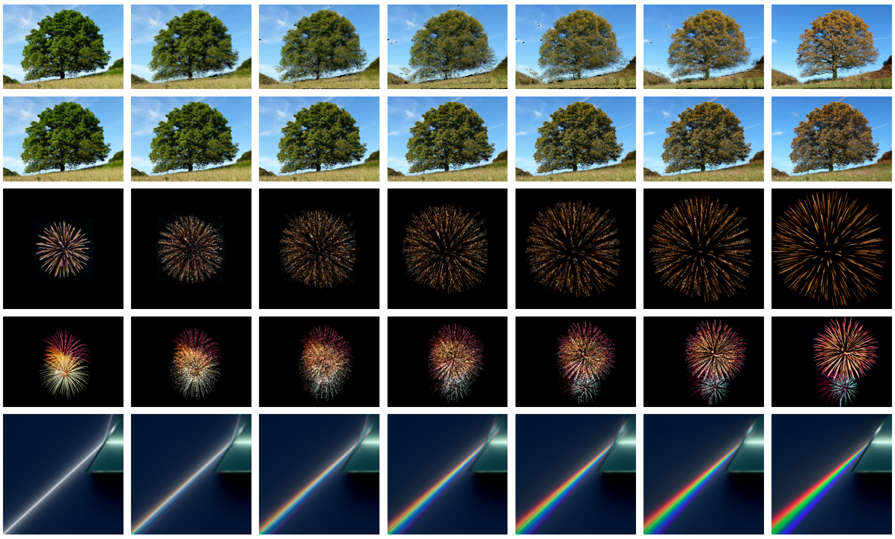

Fig. 4. Transport of 'real' images. *First two rows:* Transport between two photos of trees in different seasons using the rgb-cube model for the whole image and for the color only. *Third and fourth row:* Transport of firework images using the hsv-disk model. The hsv-diskmodel is advantageous here since there are large dark parts that get a small weight in this model. *Fifth row:* Transport between two images showing the spectrum of light using the rgb-triple model. Due to the decomposition of the spectrum with the prism the rgb-triple model is a natural choice for this example.

5 Conclusion

Standard transport-type distances are limited to scalar non-negative measures. We presented a lifting procedure to allow comparison of non-scalar signals. The method is generic and flexible, as illustrated by our examples: we computed interpolations of images over different color spaces and demonstrated the benefit of the lifted distances for classification in a simple machine learning example. Future work comprises the study of more complex signal manifolds, such as spheres and SPD matrices, as well as the computation of corresponding barycenters.

Acknowledgements Funding by the German Research Foundation (DFG) within the Research Training Group 1932 "Stochastic Models for Innovations in the Engineering Sciences", project area P3, is gratefully acknowledged.

References

1. Ahuja, R.K., Magnanti, T.L., Orlin., J.B.: Network Flows: Theory, Algorithms, and Applications. Prentice-Hall, Inc. (1993)
2. Ambrosio, L., Gigli, N., Savaré, G.: Gradient Flows in Metric Spaces and in the Space of Probability Measures. Springer Science & Business Media (2006)

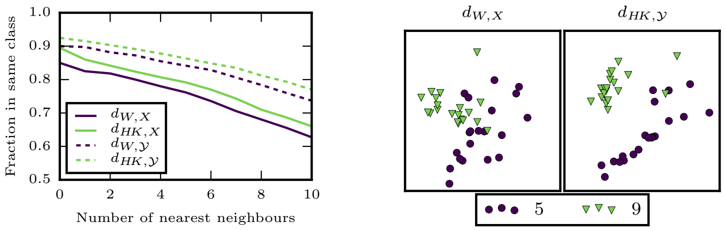

Fig. 5. *Left:* Nearest neighbour retrieval on MNIST dataset with various transport-type metrics. We randomly selected 20 samples from each class and computed the metric matrix. The plot shows the fraction of samples from the same class among the closest n neighbours w.r.t. different metrics. Using the Hellinger-Kantorovich metric improves performance relative to the Wasserstein metric; lifting the images improves performance over the standard comparison. *Right:* 2-d classical multi-dimensional scaling projection [4] of $d_{W,X}$ and $d_{HK,\mathcal{Y}}$ for classes '5' and '9'.

3. Benamou, J.D., Brenier, Y.: A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. Numerische Mathematik 84(3), 375–393 (2000)
4. Borg, I., Groenen, P.J.F.: Modern Multidimensional Scaling. Springer Series in Statistics, Springer New York, 2nd edn. (2005)
5. Chizat, L., Peyré, G., Schmitzer, B., Vialard, F.X.: An interpolating distance between optimal transport and Fisher-Rao metrics. Found. Comp. Math. (2016)
6. Chizat, L., Peyré, G., Schmitzer, B., Vialard, F.X.: Scaling algorithms for unbalanced transport problems. http://arxiv.org/abs/1607.05816 (2016)
7. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: Advances in Neural Information Processing Systems. pp. 2292–2300 (2013)
8. Fitschen, J.H., Laus, F., Steidl, G.: Transport between RGB images motivated by dynamic optimal transport. J. Math. Imaging Vis. pp. 1–21 (2016)
9. Gonzalez, R.C., Woods, R.E., Eddins, S.L.: Digital Image Processing using MATLAB, vol. 2. Gatesmark Publishing Knoxville (2009)
10. Kondratyev, S., Monsaingeon, L., Vorotnikov, D.: A new optimal transport distance on the space of finite Radon measures. http://arxiv.org/abs/1505.07746 (2015)
11. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
12. Liero, M., Mielke, A., Savaré, G.: Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures. ArXiv Preprint 1508.07941 (2015)
13. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vis. 40(2), 99–121 (2000)
14. Schmitzer, B.: Stabilized sparse scaling algorithms for entropy regularized transport problems. https://arxiv.org/abs/1610.06519 (2016)
15. Thorpe, M., Park, S., Kolouri, S., Rohde, G.K., Slepčev, D.: A transportation Lp distance for signal analysis. https://arxiv.org/abs/1609.08669 (2016)
16. Villani, C.: Optimal Transport: Old and New, Grundlehren der mathematischen Wissenschaften, vol. 338. Springer (2009)