Denoising of diffusion magnetic resonance images using a modified and differentiable Monge-Kantorovich distance for measure-valued functions

D. La Torre[*], J. Marcoux[†], F. Mendivil[‡], E.R. Vrscay[§]

July 5, 2020

Abstract

We report on the implementation of a novel total-variation denoising method for diffusion spectrum images (DSI). Our method works on the raw signal obtained from dMRI. From the Stejskal-Tanner equation [6], the signals $S(x,s_k)$, $1 \leq k \leq K$, at a given voxel location x may be considered as samplings of a measure supported on the unit sphere $\mathbb{S}^2 \in \mathbb{R}^3$ at locations $s_k = (\theta_k, \phi_k) \in \mathbb{S}^2$ which quantify the ease/difficulty of diffusion in these directions. We consider the entire signal S as a measure-valued function in a complete metric space which employs the Monge-Kantorovich (MK) metric. A total variation (TV) for measures and measure-valued functions is also defined. A major advance in this paper is the use of a modification of the standard MK distance which permits rapid computation in higher dimensions. An added bonus is that this modified metric is differentiable. The resulting TV-based denoising problem is a convex optimization problem.

2010 Mathematics Subject Classification: 94A08, 92C55, 90C08

Keywords: *total variation denoising, Monge-Kantorovich metric, measure-valued functions, diffusion MRI*

1 Introduction

In [11], a novel framework for the representation of images obtained from diffusion magnetic resonance imaging (dMRI) [8] and their denoising using total variation was

[*]SKEMA Business School - Université de la Côte-d'Azur, Sophia Antipolis Campus, France Email: davide.latorre@skema.edu

[†]Department of Mathematics and Statistics, Acadia University, Wolfville, Nova Scotia, Canada. Email: 138165m@acadiau.ca

[‡]Department of Mathematics and Statistics, Acadia University, Wolfville, Nova Scotia, Canada. Email: franklin.mendivil@acadiau.ca

[§]Department of Applied Mathematics, University of Waterloo, Ontario, Canada. Email: ervrscay@uwaterloo.ca

presented. In this framework, the (digital) images obtained from dMRI are represented by *measure-valued functions* (MVFs): If we let $X \subset R^n$ denote the *base space*, i.e., the support of an image (e.g., a human brain), then the dMR image is a mapping $\mu : X \to \mathscr{M}(Y)$, where $Y \subset \mathbb{R}^m$, $m = 1, 2$ or 3, denotes an appropriate *range space* and $\mathscr{M}(Y)$ the set of Borel probability measures on Y. For example, if $Y = \mathbb{S}^2 = \{u \in \mathbb{R}^3 \mid \|u\|_2 = 1\}$, the unit sphere in $\mathbb{R}^3$, then the measure $\mu(x) \in \mathscr{M}(\mathbb{S}^2)$ can characterize the relative ease or difficulty of diffusion of water molecules in the voxel located at $x \in X$ in the direction $s = \Omega = (\theta, \phi) \in \mathbb{S}^2$.

Let us contrast this framework with the conventional way in which images from dMRI, in particular, *high angular diffusion resolution imaging* (HARDI), are represented, namely, as *vector-valued functions*, i.e., $u(x) = (u_1(x), u_2(x), \cdots u_K(x))$, composed of signals, $u_k(x)$, which are obtained by aligning the linear *diffusion gradient vector* with gradient $g \in \mathbb{R}^3$ in K different directions, $s_k = (\theta_k, \phi_k) \in S^2$, $1 \leq k \leq K$. Most image processing algorithms simply work with such vector-valued functions in a standard manner, treating them as "cubes". What we are proposing is to work with dMRI signals in a manner that attempts to transcend such approaches in which the net signal u is simply regarded as a "stack" of component signals u_k.

In an even earlier paper [13], we examined the use of *function-valued mappings* (FVMs) for HARDI. In this case, a HARDI signal is represented by the function-valued mapping $u : X \to L^2(\mathbb{S}^2)$. In all of these works – including this paper – our philosophy has been, as described in our more recent paper on FVMs and their use in hyperspectral imaging [17], to investigate if "vector-valued images be better understood, and perhaps better algorithms be developed, if their range were a space of *continuously defined functions* instead of $\mathbb{R}^N$." Indeed, with continued improvements in technology, the number of components K in dMRI, as well as in hyperspectral imaging for remote sensing [1], is steadily increasing, essentially approaching the "continuum limit" of FVM and MVF.

In [11], our total variation-based denoising problem was formulated as follows: Given a noisy (measure-valued) image $\tilde{\mu}$, find a solution to the following optimization problem,

$$\min_{\nu \in \mathscr{F}(X)} d_{\mathscr{F}(X)}(\tilde{\mu}, \nu) + \|\nu\|_{TV}\,. \tag{1.1}$$

Here, $\mathscr{F}(X)$ denotes an appropriate space of measure-valued functions suppported on X, with metric $d_{\mathscr{F}}(X)$, and $\|\nu\|_{TV}$ denotes the total variation of the measure-valued function $\nu \in \mathscr{F}(X)$. Both of these will be reviewed in the next section.

As discussed in [11], the metric $d_{\mathscr{F}(X)}$ requires an appropriate metric between measures in the space $\mathscr{M}(Y)$. A natural choice is the so-called *Monge-Kantorovich (MK) metric* for measures, also to be reviewed below. Unfortunately, the computation of the traditional MK distance between (discrete) probability measures is computationally inexpensive only in one dimension, i.e., measures on $\mathbb{R}$. The determination of efficient algorithms to compute MK distances in $\mathbb{R}^2$ or $\mathbb{S}^2$ and $\mathbb{R}^3$ is still an open problem. In [11], we showed how to solve the optimization problem of denoising using total variation minimization for the rather artificial – but still mathematically interesting – one-dimensional case, exploiting the fact that the MK distance between two one-dimensional measures is the L^1 distance between their respective cumulative distribution functions. For the much more difficult higher-dimensional case, as we wrote

in [11], "an alternative is to replace the Monge-Kantorovich metric on measures with another metric." This is indeed one of the major accomplishments of this paper, representing a major advance over [11]: We propose a modification of the MK metric which is computationally inexpensive not only in one dimension but also in higher-dimensional spaces. An added bonus is that this modified metric is differentiable. The resulting total variation denoising scheme for measure-valued functions becomes a convex optimization problem which can be applied to real data, as we show below.

In closing this section, we review a few basic ideas from dMRI which provide the basis of our measure-valued approach. The reader may recall that in standard magnetic resonance imaging (MRI), a linear three-dimensional magnetic field is used so that the MRI signal is a Fourier transform. Inversion of this signal produces an "image" of the object being examined. In dMRI, a linear magnetic field is also used to define the direction of diffusion which is being examined. By means of an appropriate pulsing of a radio-frequency field in a *spin-echo measurement*, the relative ease/difficulty of the diffusion of "active" molecules, principally water, is measured. The strength of the signal obtained at a voxel located at $x \in X$ in the direction $s \in \mathbb{S}^2$ is given by the followng form of the so-called Stejskal-Tanner equation [8],

$$S(x,s) = S_0(x) \exp\left(-b\hat{g}_s^T D(x) \hat{g}_s\right) , \tag{1.2}$$

where

- $S_0(x)$ is the strength of the signal obtained in the absence of diffusion,
- $b > 0$ is the so-called *b-value* which is determined by the physics behind the measurement, i.e., the gyromagnetic ratio, γ, along with the pulsing time T_1,
- $\hat{g}_s$ is the unit vector with direction $s \in \mathbb{S}^2$, the direction of the gradient of the linear diffusion gradient field,
- $D(x)$ is the 3×3 *diffusion tensor* at x. If the medium is isotropic in a neighbourhood of x, then $D = dI$, where d is a constant (the *diffusion coefficient*) and I is the identity matrix.

In HARDI imaging, the signal attenuation can also be modelled as follows [9],

$$S(x,s) = S_0(x) \exp((-bd(x,s)) , \tag{1.3}$$

where $d(x,s)$, the *spherical Apparent Diffusion Coefficient* (sADC), characterizes the rate of diffusion in the direction $s \in \mathbb{S}^2$. The sADC $d(x,s)$ may be viewed as a function defined on the unit sphere $\mathbb{S}^2$, which provides a local map of the rates of diffusion of water molecules from location x in all (or at least experimentally realizable) directions. A related quantity is the *diffusion orientation probability distribution* (dODF) $O(x,s)$, which is defined to be the probability that a diffusing water molecule at x moves in direction s [8]. (As is well known – and we simply mention it here to keep our discussion brief – the method of *tractography* [2, 12] uses this information to determine the anatomical structure of white matter in the brain by means of *connectomes*, visual representations of neural fibre connections. This information can then be used for diagnostic purposes [4].)

In this study, we choose to work with the raw experimental data $S(x,s_k)$, $1 \leq k \leq K$, which is contaminated – usually quite highly – with instrument noise. Our goal is to denoise this data. From Eqs. (1.2) and (1.3), we see that diffusion attenuates the (theoretically noiseless) signal S. For a given $x \in X$, a direction for which $d(x,s)$ (or $O(x,s)$) is large (small), i.e., a large diffusion rate, corresponds to a small (large) value of the signal strength $S(x,s)$. This implies that the signal strength $S(x,s)$ is (exponentially) inversely related to the diffusion rate at x in the direction s – essentially a measure of the *resistance* to diffusion. Even though, for each $x \in X$, the signal $S(x,s)$, considered as a function of $s \in \mathbb{S}^2$ does not define a measure over $\mathbb{S}^2$ (in contrast to the ODF defined earlier), *we propose to treat is as a probability measure* (after suitable normalization). The motivation is that *measure-based distance functions, such as the Monge-Kantorovich metric, should work better than simple metrics for functions such as L^p to characterize differences in information – in this case anisotropic diffusion – that is contained in the dMRI data*. Indeed, there has been interest in the use of metrics derived from information theory (e.g., Fisher-Rao metric, von Mises-Fisher distribution) to measure the overlap/difference between ODFs – see the discussion and references contained in [9] – but, to the best of our knowledge, no use of such metrics in any denoising algorithm has yet been performed, apart from [11] and this paper.

2 Theoretical background

2.1 Measure-valued images

As mentioned in Section 1, we let $X \subset \mathbb{R}^n$ denote our "base space," the physical space which contains the object being imaged (e.g., a human brain). Also let $Y \subset \mathbb{R}^m$ be a compact *range space* and $\mathscr{M}(Y)$ the set of probability measures on Borel subsets of Y. In our particular applications below, $m = n = 3$, with $X = [0,1]^3$ and $Y = \mathbb{S}^2$, the unit sphere in $\mathbb{R}^3$, which represents the set of all directions from any point $x \in X$.

For any two measures $\alpha, \beta \in \mathscr{M}(Y)$, the *Monge-Kantorovich distance* [10] is defined as follows,

$$d_{MK}(\alpha,\beta) = \sup\left\{ \int_Y \phi(t) d(\alpha-\beta)(t) : \phi \in \mathrm{Lip}_1(Y) \right\}, \tag{2.4}$$

where, as usual, $\mathrm{Lip}_1(Y) = \{f : Y \to \mathbb{R} : |f(x) - f(y)| \leq \|x-y\| \ \forall x,y \in Y\}$. We mention that the MK distance is well-defined as long as the two measures have the same total mass. (This implies that, in general, they do not have to be probability measures.) Furthermore, convergence of a sequence of measures $\alpha_n \in \mathscr{M}(Y)$ in the Monge-Kantorovich metric is equivalent to weak convergence of α_n. Using this fact it is not difficult to show that $(\mathscr{M}(Y), d_{MK})$ is a complete metric space.

The Monge-Kantorovich distance (which is special case of the Wasserstein metric [21]) has its origins in the theory of mass transport and is the solution to the dual of a linear programming formulation of the mass transportation problem. Because of this beginning, the MK distance has many useful geometric properties which make it a natural "extension" of the underlying metric on Y to the set of measures on Y. For example, $d_{MK}(\delta_x, \delta_y) = \|x-y\|$ for point masses δ_x and δ_y in $\mathbb{R}^d$ situated at x and y,

respectively (it is straightforward to show that $\phi(t) = d(x,t)$ is an optimal function to use in the definition of the MK distance given in (2.4)) Additionally, for the special case of two compactly supported probability measures α, β on $\mathbb{R}$, we have (see [20])

$$d_{MK}(\alpha,\beta) = \int_{\mathbb{R}} |\alpha(-\infty,t] - \beta(-\infty,t]|\, dt,$$

(the L^1 norm of the difference between their respective cumulative distribution functions). It is this connection between d_{MK} on $\mathscr{M}(Y)$ and d on Y that makes the MK distance a good choice in any problem where the geometry of their supports is a significant factor when comparing two measures.

The set of measure-valued images used in our framework is defined as follows,

$$\mathscr{F}(X) = \{\mu : X \to \mathscr{M}(Y) \mid x \mapsto \mu(x)(A) \text{ is measurable } \forall \text{ Borel subsets } A \subseteq Y\}. \quad (2.5)$$

There are many different and natural choices for a metric on $\mathscr{F}(X)$. We will use the following,

$$d_{\mathscr{F}(X)}(\mu,\nu) = \left(\int_X d_{MK}(\mu(x),\nu(x))^2\, dx\right)^{1/2}, \quad \mu,\nu \in \mathscr{F}(X). \quad (2.6)$$

Because Y is a compact metric space, we have that $\mathscr{M}(Y)$ is compact under the Monge-Kantorovich distance and therefore complete. This implies that $\mathscr{F}(X)$ is complete when we use (2.6) as the metric. In addition, $\mathscr{M}(Y)$ is convex, which implies that $\mathscr{F}(X)$ is complex. This, in turn, implies that $\mathscr{F}(X)$ is a closed convex set. This will be important for us in the sequel.

2.2 Total variation

The basic idea behind total variation denoising is that adding noise causes the total variation to increase, so if one can decrease the total variation of an image in a controlled way this noise can be reduced. Classical total variation regularization procedures have been found to work very well in removing unwanted detail while still preserving edges [19]. For a differentiable function $f : A \subset \mathbb{R}^n \to \mathbb{R}$, the standard total variation is defined as follows,

$$\|f\|_{TV} = \int_X \|\nabla f\|_2\, dx. \quad (2.7)$$

Many other variants of total variation exist – see [7] for an overview of some recent ones.

In [11] we introduced the following notion of total variation for measure-valued images $\mu \in \mathscr{F}(X)$,

$$\|\mu\|_{TV,MK} = \int_X \|D\mu\|_2\, dx = \int_X \left(\sum_{i=1}^n |D_i\mu(x)|^2\right)^{1/2} dx, \quad (2.8)$$

where

$$|D_i\mu(x)| := \sup_{\phi_i \in \mathrm{Lip}_1(Y)} \limsup_{h_i \to 0^+} \frac{1}{h_i} \int_{t\in Y} \phi_i(t)\, d(\mu(x + \hat{e}_i h_i) - \mu(x))(t), \quad 1 \le i \le 3, \quad (2.9)$$

are the analogues of the magnitudes of the directional derivatives of μ at the point $x \in X$. Since $\mu(x)$ is a probability measure on Y for each x, we have that

$$\left| \int_{t \in Y} \phi_i(t)\, d(\mu(x + \hat{e}_i h_i) - \mu(x))(t) \right| \leq 2\,\mathrm{diam}(Y)$$

for all x. This means that Eq. (2.8) mainly measures the oscillations of μ as a function of x (since μ cannot become "unbounded").

In the computation of the magnitudes in (2.9), the supremum over $\phi_i \in \mathrm{Lip}_1(Y)$ arises from the use of the Monge-Kantorovich metric on measures. (As a point of interest, the use of the *total variation norm* on measures would necessitate the use of a supremum over $\phi_i \in L^\infty(Y)$.) If $\mu(x) = \delta_{f(x)}$ for some differentiable $f : X \to \mathbb{R}$, then one can show that

$$\|\mu\|_{TV,MK} = \int_X \|\nabla f\|_2 \, dx$$

and so our definition reduces to the classical one.

If $\mu(x)$ has a density $\rho_\mu(x,\cdot)$ for each $x \in X$, then a standard calculation shows that $D_i\mu$ is a signed measure with density $\frac{\partial \rho_\mu}{\partial x_i}$. This is very useful for models where one fits data with a parametric form of $\mu(x)$ for each x.

Our total variation-based denoising algorithm for measure-valued images in $\mathscr{F}(X)$ is then given as follows: Given a noisy image $\tilde{\mu}$ (the "observed data") we seek a solution to

$$\min_{\nu \in \mathscr{F}(X)} \Psi(\nu) := \min_{\nu \in \mathscr{F}(X)} \left\{ (1-\lambda) \left(\int_X d_{MK}(\tilde{\mu}(x), \nu(x))^2 \, dx \right)^{1/2} + \lambda \|\nu\|_{TV,MK} \right\}, \tag{2.10}$$

where $0 < \lambda < 1$ is the regularization parameter. Since both terms in the objective function in (2.10) are derived from norms on $\mathscr{F}(X)$ the overall objective function is also a convex function of ν. Since $\mathscr{F}(X)$ is convex, this means that there is always a solution to (2.10) for any λ.

3 Discretization

Our numerical experiments were performed using data from the Stanford Digital Repository (https://purl.stanford.edu/ng782rw8378) – see also [18]. This data was acquired from two human subjects and was measured using 150 different directions (the directions are shown in Figure 1). Our discussion of the discretization below is specific to this data set but the generalization to another is clear.

For the purposes of computation the domain space of the image, X, is partitioned into a 3D grid of *voxels* and the measure-valued function μ gives a (fixed) probability distribution for each such voxel. Our voxel grid is $81 \times 106 \times 76$. In addition, the directions in which diffusion is measured are represented by points on the unit sphere (which is acting as the support, Y, of the measure $\mu(x)$). Since the data was measured in 150 different directions, we take a 150-point discretization, $\mathscr{G}$, of the unit sphere (shown in Figure 1) and assign a weight, $w_{i,j}$, to each edge (i,j) of this graph according

to the distance (on the sphere) between the corresponding points. Thus at each voxel, x, our measure-valued function μ is specified by a probability vector of length 150.

The distance between two distinct points $p, q \in \mathscr{G}$ is defined as the minimum total weight of any path between p and q. Because the weights of the edges of $\mathscr{G}$ are obtained by distances in $\mathbb{R}^3$, the triangle inequality implies that the distance between adjacent vertices is just the weight of the edge between them (as we would expect).

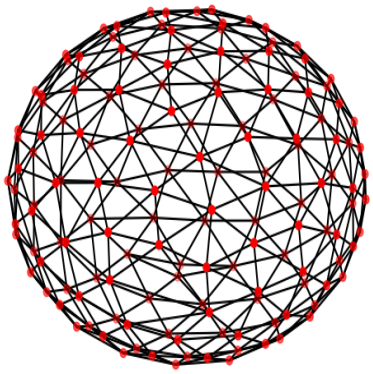

Figure 1: 150-point discretization of the sphere

As presented in [14] (see also [15] for another nice viewpoint) a natural way to interpret the Monge-Kantorovich metric for a finite metric space is to model the metric space by a graph and use a difference operator as the analogue of the derivative. Recall that for a graph $\mathscr{G}$, the *vertex space*, $V(\mathscr{G})$, and *edge space*, $E(\mathscr{G})$, are vectors spaces of formal linear combinations of vertices and edges, respectively.

The *weighted edge-incidence* matrix D of $\mathscr{G}$ has rows corresponding to the edges and columns corresponding to the vertices of $\mathscr{G}$. After orienting the edges of $\mathscr{G}$ (with any orientation), the row corresponding to $i \to j$ has a $-w_{i,j}$ in column i and a $+w_{i,j}$ in column j of the edge-incidence matrix. The kernel of D is easily seen to consist of constant vectors in $V(\mathscr{G})$. Then the MK distance between α and β in $\mathscr{M}(\mathscr{G})$ is (see [14])

$$d_{MK}(\alpha,\beta) := \sup\{f \cdot (\alpha - \beta) : f \in V(\mathscr{G}), \|Df\|_\infty \leq 1\}. \tag{3.11}$$

The condition $\|Df\|_\infty$ is the analogue of $f \in \mathrm{Lip}_1(Y)$ and it is this condition which underlies the difficulty in computing (3.11).

Modified Monge-Kantorovich metric

The use of the infinity norm causes differentiability problems and so we use a modified version of (3.11)

$$d_{MK,2}(\alpha,\beta) = \sup\{f \cdot (\alpha - \beta) : f \in V(\mathscr{G}), \|Df\|_2 \leq 1\}. \tag{3.12}$$

As usual, the use of the Euclidean norm is used to ensure differentiability. However as we will see next, this also allows one to obtain a simple formula for this distance.

As a point of notation, we use $A^\dagger$ to denote the pseudo-inverse (or Moore-Penrose inverse) of a matrix A (see [3]) and is defined and unique for any matrix A (regardless of size). The defining properties of $A^\dagger$ are

1. $AA^\dagger A = A$,
2. $A^\dagger AA^\dagger = A$,
3. $(AA^\dagger)^T = AA^\dagger$, and
4. $(A^\dagger A)^T = A^\dagger A$.

In particular, these properties imply that $AA^\dagger$ is the orthogonal projection onto range(A) and $A^\dagger A$ is the orthogonal projection onto $\ker(A)^\perp$ and so $\|AA^\dagger\|_2 = \|A^\dagger A\|_2 = 1$. In addition, $\ker(A^\dagger)^\perp = \text{range}(A)$ and $\ker(A)^\perp = \text{range}(A^\dagger)$.

Proposition 3.1. *Let $\mathscr{G}$ be a graph with D its weighted edge-adjacency matrix (with an arbitrary orientation for the edges). Then for any two probability distributions α, β on the vertices of $\mathscr{G}$, we have*

$$\sup\{f \cdot (\alpha - \beta) : \|Df\|_2 \leq 1\} = \|(D^\dagger)^T(\alpha - \beta)\|_2.$$

Proof. First we show that

$$\sup\{(\alpha-\beta)\cdot f : f \in V(\mathscr{G}), \|Df\|_2 \leq 1\} = \sup\{(\alpha-\beta)\cdot(D^\dagger g) : g \in E(\mathscr{G}), \|DD^\dagger g\|_2 \leq 1\}. \tag{3.13}$$

(This equality is actually independent of the norm we use ($\|\cdot\|_2$ in this case)). One direction is a simple consequence of the fact that $g \in E(\mathscr{G})$ implies that $D^\dagger g \in V(\mathscr{G})$ and thus

$$\sup\{(\alpha-\beta)\cdot f : f \in V(\mathscr{G}), \|Df\|_2 \leq 1\} \geq \sup\{(\alpha-\beta)\cdot(D^\dagger g) : g \in E(\mathscr{G}), \|DD^\dagger g\|_2 \leq 1\}.$$

For the converse, we first note that $\ker(D) \subset V(\mathscr{G})$ is the set of constant vectors and so $\alpha - \beta \in \ker(D)^\perp$. Next we note that $D^\dagger D$ is the orthogonal projection onto $\ker(D)^\perp$ and thus $f - D^\dagger Df \in (\ker(D)^\perp)^\perp = \ker(D)$. This means $(\alpha-\beta)\cdot(f - D^\dagger Df) = 0$ or $(\alpha-\beta)\cdot f = (\alpha-\beta)\cdot(D^\dagger Df)$. Thus if $f \in V(\mathscr{G})$ setting $g = Df$ gives $(\alpha-\beta)\cdot(D^\dagger g) = (\alpha-\beta)\cdot f$ and $\|Df\|_2 \leq 1$ implies that $\|DD^\dagger g\|_2 \leq \|DD^\dagger\|_2\|Df\|_2 \leq 1$ as well. This gives the other direction.

Our next step is to show that

$$\sup\{(\alpha-\beta)\cdot(D^\dagger g) : g \in E(\mathscr{G}), \|DD^\dagger g\|_2 \leq 1\} = \sup\{(\alpha-\beta)\cdot(D^\dagger g) : g \in E(\mathscr{G}), \|g\|_2 \leq 1\}. \tag{3.14}$$

Again one direction (the $\leq$ direction) is simple, this time being the consequence of $\|DD^\dagger\|_2 = 1$. For the other direction, we note that since $D^\dagger DD^\dagger = D^\dagger$, we have

$$D^\dagger(g - DD^\dagger g) = D^\dagger g - D^\dagger DD^\dagger g = 0.$$

Thus if $\|DD^\dagger g\|_2 \leq 1$, we set $h = DD^\dagger g$ so $\|h\|_2 \leq 1$ and $(\alpha-\beta)\cdot(D^\dagger g) = (\alpha-\beta)\cdot(D^\dagger h)$, which gives the other inequality.

Finally, with these two equalities we have

$$\begin{aligned}\sup\{(\alpha-\beta)\cdot f : f \in V(\mathscr{G}), \|Df\|_2 \leq 1\} &= \sup\{(\alpha-\beta)\cdot (D^\dagger g) : g \in V(\mathscr{G}), \|g\|_2 \leq 1\} \\ &= \sup\{(D^\dagger)^T(\alpha-\beta)\cdot g : g \in V(\mathscr{G}), \|g\|_2 \leq 1\} \\ &= \|(D^\dagger)^T(\alpha-\beta)\|_2,\end{aligned}$$

as desired. □

As a result, we will use the metric

$$d_{MK,2}(\alpha,\beta) = \|(D^\dagger)^T(\alpha-\beta)\|_2, \tag{3.15}$$

for a modified and differentiable version of the discrete MK distance. Notice that (3.15) is no longer a linear programming problem but a simple computation. Using this metric, our discretized version of the objective function in (2.10) is given by

$$\left\{(1-\lambda)\left(\sum_{i\in\text{voxels}} d_{MK,2}(\tilde{\mu}(i),\nu(i))^2\,dx\right)^{1/2} + \lambda\|\nu\|_{TV,MK}\right\}. \tag{3.16}$$

In addition, our discretized version of (2.10) is also convex, so it can be optimized by many standard methods.

4 Numerical experiments

We performed (using a mixture of MATLAB and python) preliminary numerical experiments both with small images and large images. The small images were extracted from the large ones and were used extensively in tuning the optimization process (in particular the weight λ).

For our preliminary experiments we used a variant of gradient descent with a careful control of the step size. Since we are using a modified Monge-Kantorovich distance, it is important that the total mass at each voxel remains constant throughout the algorithm (otherwise the MK distance is not defined). Conveniently the gradient of the objective function in (2.10) is a zero-sum vector (as we show below) for each voxel and thus when we subtract any multiple of the gradient from the current state (dMRI image) of the process the total mass at each voxel is preserved.

The step size control functioned both to ensure a decreasing objective function (as is usual with step size control for gradient descent) but also to ensure that none of the components at any voxel became negative (and thus no longer represented a probability distribution). Specifically, if we are using

$$\nu_{n+1} = \nu_n - \tau\nabla\Psi$$

(where Ψ is the objective function from (3.16)) we must insure that no component of ν_{n+1} is negative by choosing a small enough value for the multiplier τ. However, in practice very quickly this process requires the value of τ to be smaller than 10^{-9} and so the steps were too small to influence the objective function appreciably.

In order to deal with this issue we implemented two changes to the standard gradient descent. Both of these modify $\nabla\Psi$, but make different "local" modifications at different voxels. The first method is to allow the value of τ to depend on the voxel. We implemented this change by scaling $\nabla\Psi$ differently at each voxel. Having done this, if the required scaling τ_i at voxel i is "too small', we simply set $\nabla\Psi$ restricted to this voxel to be zero. Using these two methods we obtain $\widehat{\nabla\Psi}$, a modification to the actual gradient. It is not hard to see that $\nabla\Psi \cdot \widehat{\nabla\Psi} \geq 0$ and so $-\widehat{\nabla\Psi}$ is also a "downhill" direction for Ψ.

Each image has about 98 million data values and so the optimization algorithm runs slowly (about two iterations per hour on a workstation). We discuss the details of the objective function in the next section.

Discretized objective function

We now give some details about the two parts of the discrete version of (2.10) as given in (3.16). The first term of the objective function is

$$\left(\sum_I \|(D^\dagger)^T(\tilde{\mu}_I - \nu_I)\|_2^2\right)^{1/2} = \left(\sum_I (\tilde{\mu}_I - \nu_I)^T D^\dagger (D^\dagger)^T (\tilde{\mu}_I - \nu_I)\right)^{1/2}, \qquad (4.17)$$

where the sum is over all the voxels I. The second part of the objective function is written in terms of the magnitudes of the directional derivatives in the three spatial directions,

$$\left(\sum_I \|D^{\dagger T}(\nu_I - \nu_{I'})\|_2^2 + \|D^{\dagger T}(\nu_I - \nu_{I''})\|_2^2 + \|D^{\dagger T}(\nu_I - \nu_{I'''})\|_2^2\right)^{1/2}. \qquad (4.18)$$

To understand this equation, we have to see how we compute the total variation. For a function $f : \mathbb{R}^3 \to \mathbb{R}$ one can use something like $\int \|\nabla f(x)\|\, dx$ so we will use a finite-difference approximation to the spatial directional derivatives. For a voxel $I = (a,b,c)$ let the voxel $I' = (a-1,b,c)$ and $I'' = (a,b-1,c)$ and $I''' = (a,b,c-1)$). Then a very rough approximation to the gradient of a measure-valued function μ is $\langle \mu_I - \mu_{I'}, \mu_I - \mu_{I''}, \mu_I - \mu_{I'''} \rangle$. We compute the MK norm of this difference at each voxel.

Now we will show that the gradient of our objective function is a zero-sum vector for each voxel. To do this we first find expressions for the partial derivatives of the objective function. Let ξ be some parameter on which ν depends (e.g., one component

of v). For the first term (4.17) in the objective function, we have

$$\begin{aligned}
&\frac{\partial}{\partial \xi}\left(\sum_I \|(D^\dagger)^T(\tilde{\mu}_I - v_I)\|_2^2\right)^{1/2} \\
&= \frac{1}{2}\left(\sum_I (\tilde{\mu}_I - v_I)^T D^\dagger (D^\dagger)^T (\tilde{\mu}_I - v_I)\right)^{-1/2}\left[\sum_I 2\frac{\partial}{\partial \xi}(\tilde{\mu}_I - v_I)^T D^\dagger (D^\dagger)^T (\tilde{\mu}_I - v_I)\right] \\
&= \frac{-\sum_I \left(\frac{\partial v_I}{\partial \xi}\right)^T D^\dagger (D^\dagger)^T (\tilde{\mu}_I - v_I)}{\sqrt{\sum_I \|D^{\dagger T}(\tilde{\mu}_I - v_I)\|_2^2}}. \qquad (4.19)
\end{aligned}$$

For the gradient we take ξ to be one component of v at one voxel, say component i at voxel I. Doing this causes $\frac{\partial v_I}{\partial \xi}$ to be a vector with 150 components all of which are zero except with a 1 in component i. Thus when we multiply $D^\dagger(D^\dagger)^T(\tilde{\mu}_I - v_I)$ by $\frac{\partial v_I}{\partial \xi}$ we are simply extracting the ith component of $D^\dagger(D^\dagger)^T(\tilde{\mu}_I - v_I)$. When done for all components i, the result is simply the vector $D^\dagger(D^\dagger)^T(\tilde{\mu}_I - v_I)$ itself. Now, $\text{range}(D^\dagger) = \ker(D)^\perp \subset V(\mathscr{G})$ is the subspace of zero-sum vectors and thus $D^\dagger(D^\dagger)^T(\tilde{\mu}_I - v_I)$ is a zero-sum vector. Since this is true for each voxel I, this shows that the gradient of the first term (4.17) is zero-sum for each voxel.

Moving to the second term (4.18), we have

$$\begin{aligned}
\frac{\partial}{\partial \xi}\|v\|_{TV,MK} &= \frac{\sum_I \sum_{J=I',I'',I'''} \frac{\partial}{\partial \xi}\left((v_I - v_J)^T D^\dagger (D^\dagger)^T (v_I - v_J)\right)}{2\left(\sum_I \|(D^\dagger)^T(v_I - v_{I'})\|_2^2 + \|(D^\dagger)^T(v_I - v_{I''})\|_2^2 + \|(D^\dagger)^T(v_I - v_{I'''})\|_2^2\right)^{1/2}} \\
&= \frac{\sum_I \sum_{J=I',I'',I'''} \left(\frac{\partial v_I}{\partial \xi} - \frac{\partial v_J}{\partial \xi}\right)^T D^\dagger (D^\dagger)^T (v_I - v_J)}{\|v\|_{TV,MK}}. \qquad (4.20)
\end{aligned}$$

For a given voxel I, there are the six parts

$$\left(\frac{\partial v_I}{\partial \xi}\right)^T D^\dagger (D^\dagger)^T (v_I - v_J) \quad \text{and} \quad -\left(\frac{\partial v_J}{\partial \xi}\right)^T D^\dagger (D^\dagger)^T (v_I - v_J) \text{ for } J = I', I'', I''.$$

The same argument used for the first term of the objective function shows that each of these results in a zero-sum vector and thus their sum will as well. Since this is true for each voxel, it is true for the entire gradient as well. Thus for any λ the gradient of our objective function (3.16) is a zero-sum vector at each voxel I.

Numerical results

Figure 2 shows a typical run of 1000 iterations for a small image. The top plot shows how the objective function decreases (as it should) while the bottom plot shows the evolution of the distance to the original noise-free image. As expected, even the original noise-free image is transformed by the denoising algorithm.

Figure 3 shows a representative "slice" through the image data for one of the "large" images. A plane of voxels was chosen and then for each voxel in this plane the value of one component of the 150-component probability distribution is shown. The images have been contrast enhanced since the original data values are all very close to zero. In these images only 40 iterations of gradient descent were performed.

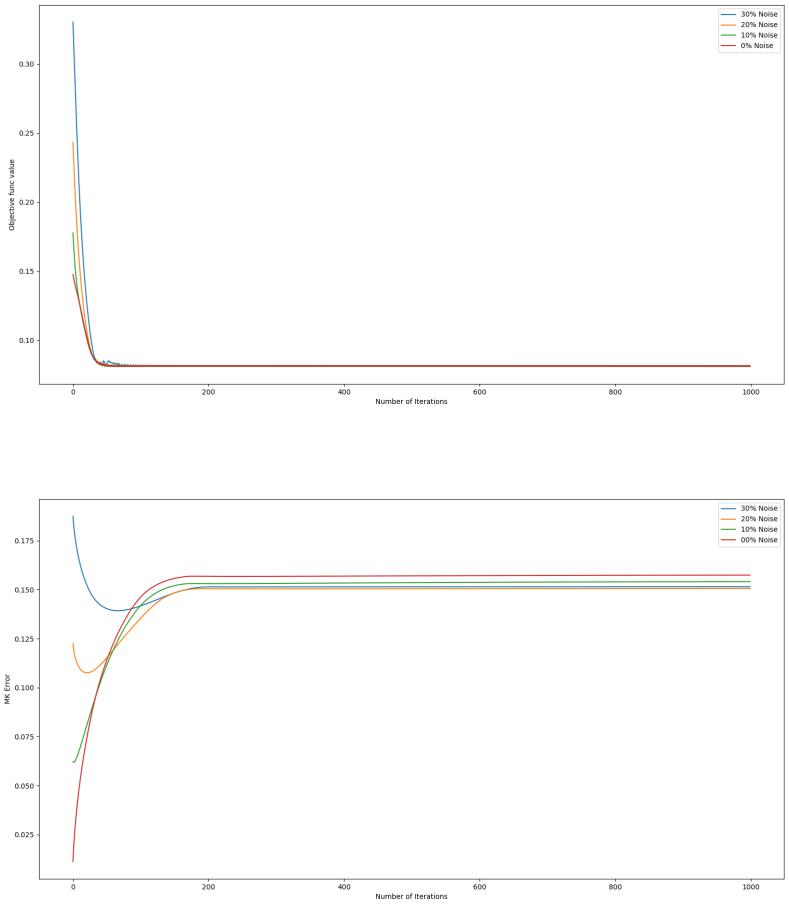

Figure 2: Convergence for small image: objective (top), error (bottom)

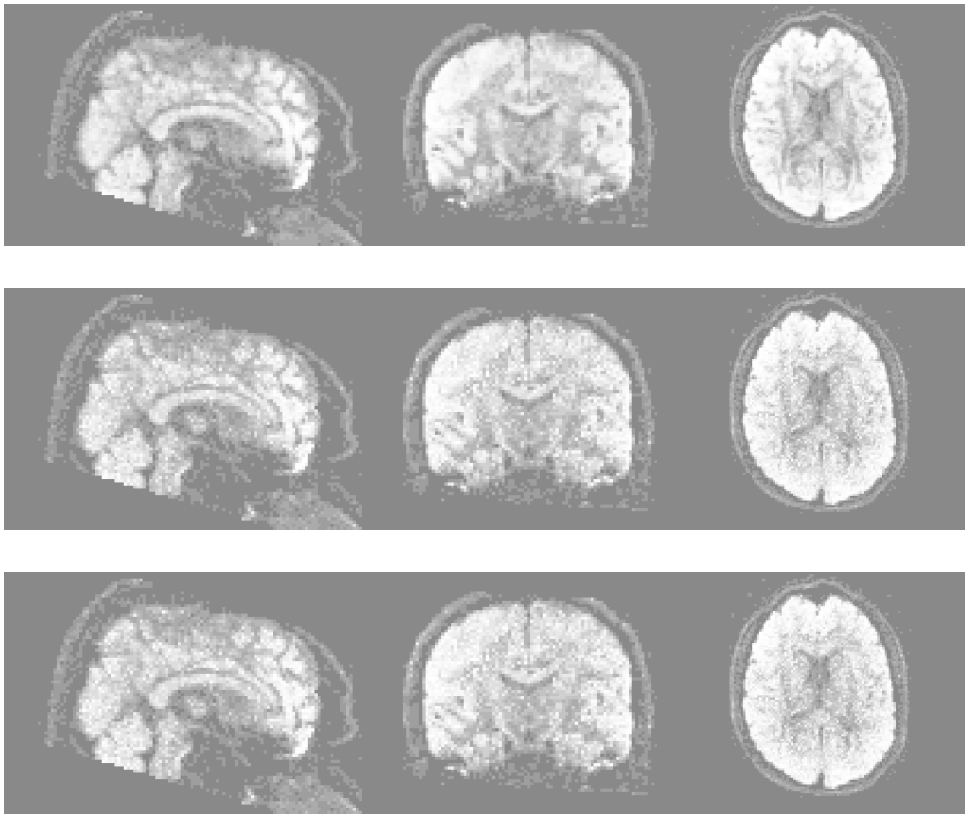

Figure 3: Representative slice through the image: original (top), 10% noise (middle), denoised (bottom)

5 Concluding Remarks

In this paper, we consider the raw signal obtained from dMRI as defining a measure-valued function. From the Stejskal-Tanner equation, at each voxel $x \in X$, the signals $S(x, s_k)$, $1 \leq k \leq K$, are considered to be samplings of a measure supported on the unit sphere $\mathbb{S}^2$. This measure characterizes the ease/difficulty of diffusion of water molecules from x. A modified version of the classical Monge-Kantorovich metric – a metric between measures – is employed in an effort to characterize the differences in information in a better way than standard L^p-based metrics. The total variation of a measure-valued function is also defined. This mathematical framework sets up a TV-based denoising algorithm which is a convex optimization problem. Some preliminary results to raw data have been presented.

Acknowledgements

This research was supported in part by Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) (ERV and FM). Financial support from Acadia University (JM) is also gratefully acknowledged.

References

[1] Airborne Visible/Infrared Imaging Spectrometer, NASA Jet Propulsion Laboratory, California Institute of Technology site: https://aviris.jpl.nasa.gov

[2] P.J. Basser, J. Mattiello, D. Le Bihan, MR diffusion tensor spectroscopy and imaging. Biophys. J. **66**(1) 259–267 (1994).

[3] A. Ben-Israel and T. Greville, Generalized inverses: Theory and applications, 2nd ed, CMS Books in Mathematics, Vol 15, Springer-Verlag, New York (2003).

[4] D. Le Bihan, E. Breton, D. Lallemand, P. Grenier, E. Cabanis, M. Laval-Jeantet, MR imaging of intravoxel incoherent motions: Application to diffusion and perfusion in neurological disorders. Radiology **161** 401–407 (1986).

[5] V. Brion, C. Poupon, O. Ri, S. Aja-Fernandez, A. Tristan-Vega, J. F. Mangin, D. Le Bihan, F. Poupon, Noise correction for HARDI and HYDI data obtained with multi-channel coils and sum of squares reconstruction: An anisotropic extension of the LMMSE. *Magnetic Resonance Imaging*, **31**, 1360–71 (2013).

[6] P.T. Callahan, *Principles of nuclear magnetic resonance microscopy*, Clarendon Press, Oxford, UK (1991).

[7] B. Goldluecke, E. Strekalovskiy, D. Cremers, The Natural Vectorial Total Variation Which Arises from Geometric Measure Theory, SIAM J. Imaging Sci., **5**(2), 537-563 (2012).

[8] H. Johansen-Berg and T.E.J. Behrens, *Diffusion MRI: From Quantitative Measurements to In-Vivo Neuroanatomy*, 1st Ed. Academic, New York (2009).

[9] Y. Kim, P.M. Thompson and L.A. Vese, HARDI data denoising using vectorial total variation and logarithmic barrier, Inverse Prob. Imaging **4**(2) 273-310 (2010).

[10] H. Kunze, D. La Torre, F. Mendivil and E.R. Vrscay, *Fractal-based methods in analysis*, Springer Verlag (2012).

[11] D. La Torre, O. Michailovich, F. Mendivil and E.R. Vrscay, Total Variation Minimization for Measure-Valued Images with Diffusion Spectrum Imaging as Motivation, in *Image Analysis and Recognition*, Proceedings of ICIAR 2016, LNCS **9730**, 131-137 (2016).

[12] J. Malcolm, M. Shenton, and Y. Rathi , Neural tractography using an unscented Kalman filter, in J.L. Prince, D.L. Pham, K.J. Myers, K.J. (Eds.) IPMI 2009. LNCS, **5636**, 126-138. Springer, Heidelberg (2009)

[13] O. Michailovich, D. La Torre and E.R. Vrscay, Function-Valued Mappings, Total Variation and Compressed Sensing for Diffusion MRI, in *Image Analysis and Recognition*, Proceedings of ICIAR 2012, LNCS **7325**, 286-295 (2012).

[14] F. Mendivil, Computing the Monge-Kantorovich distance, *Comp. Appl. Math.* **36** (3), 1389-1402 (2017).

[15] L. Montrucchio and G. Pistone, Kantorovich distance on a finite metric space, *arXiv preprint arXiv:1905.07547v5*, 2019.

[16] V. Brion, C. Poupon, O. Ri, S. Aja-Fernandez, A. Tristan-Vega, J. F. Mangin, D. Le Bihan, F. Poupon (2013) Noise correction for HARDI and HYDI data obtained with multi-channel coils and sum of squares reconstruction: An anisotropic extension of the LMMSE. *Magnetic Resonance Imaging*, **31**, 1360–71 (2013).

[17] D. Otero, D. La Torre, O. Michailovich and E.R. Vrscay, On the theory of function-valued mappings and its application to the processing of hyperspectral images, Signal Proc. **134** 185-196 (2017).

[18] A. Rokem, J.D. Yeatman, F. Pestilli, K.N. Kay, A. Mezer, S. van der Walt, B.A. Wandell, Evaluating the accuracy of diffusion MRI models in white matter, PLoS ONE **10**(4): e0 123272. doi: 10.137 1/journal.pone.0123272.

[19] D. Strong, T. Chan, Edge-preserving and scale-dependent properties of total variation regularization, *Inverse Problems* **19** 165-187 (2003).

[20] S. S. Vallender, Calculation of the Wasserstein distance between probability distributions on the line, *Theory Probab. Appl.* **18**, 784-786 (1973).

[21] C. Villani, *Optimal transport: old and new*, Springer, Berline Heidelberg (2008).