

An Efficient Algorithm for Matrix-Valued and Vector-Valued Optimal Mass Transport

Yongxin Chen¹ · Eldad Haber² · Kaoru Yamamoto³ · Tryphon T. Georgiou⁴ · Allen Tannenbaum⁵

Received: 21 February 2018 / Revised: 6 March 2018 / Accepted: 9 March 2018 /
Published online: 16 March 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract We present an efficient algorithm for recent generalizations of optimal mass transport theory to matrix-valued and vector-valued densities. These generalizations lead to several applications including diffusion tensor imaging, color image processing, and multi-modality imaging. The algorithm is based on sequential quadratic programming. By approximating the Hessian of the cost and solving each iteration in an inexact manner, we are able to solve each iteration with relatively low cost while still maintaining a fast convergence rate. The core of the algorithm is solving a weighted Poisson equation, where different efficient preconditioners may be employed. We utilize incomplete Cholesky factorization, which yields an efficient and straightforward solver for our problem. Several illustrative examples are presented for both the matrix and vector-valued cases.

Kaoru Yamamoto is a member of the LCCC Linnaeus Center and the ELLIIT Excellence Center at Lund University.

✉ Allen Tannenbaum
allen.tannenbaum@stonybrook.edu

Yongxin Chen
yongchen@iastate.edu

Eldad Haber
ehaber@eoas.ubc.ca

Kaoru Yamamoto
k.yamamoto@ieee.org

Tryphon T. Georgiou
tryphon@uci.edu

¹ Department of Electrical and Computer Engineering, Iowa State University, Ames, IA, USA

² Department of Mathematics, University British Columbia, Vancouver, Canada

³ Department of Automatic Control, Lund University, Lund, Sweden

⁴ Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA, USA

⁵ Departments of Computer Science and Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY, USA

Keywords Optimal mass transport · Vector-valued data · Matrix-valued data · Quantum mechanics

1 Introduction

The theory of optimal mass transport (OMT) [1, 31, 36] has proven its power and usefulness in both theory and applications. The theory part has been developed through a sequence of elegant papers, and the research is still going strong; see [4, 7, 17, 19, 20, 25, 27, 30] and the references therein. On the other hand, during the past decade, the need for applications has engendered the fast development of efficient algorithms for OMT [2, 5, 6, 10, 15, 18, 22, 24]. Recently, the OMT theory has been extended to study matrix-valued [9, 11, 26] and vector-valued densities [12].

The mathematical approach to matrix optimal mass transport in [9, 11, 26] is based on the seminal work of Benamou-Brenier [4], where optimal mass transport with quadratic cost is recast as the problem of minimizing kinetic energy (i.e., an action integral) subject to a continuity equation. In the matrix case, one needs to develop a non-commutative counterpart to scalar optimal transport where probability distributions are replaced by density matrices ρ (Hermitian positive-definite with unit trace) and where “transport” corresponds to a flow on the space of such matrices that minimizes a corresponding action integral. The work is motivated by a plethora of applications including spectral analysis of vector-valued time-series, which may encode different modalities (e.g., frequency, color, polarization) across a distributed array of sensors [28]. The associated power spectra are matrix-valued and hence there is a need for suitable metrics that quantify distances and provide tools to average and interpolate spectra. The generalization of the Benamou-Brenier theory is founded upon concepts from quantum mechanics, and allows us to formulate a continuity equation for matrix-flows, and then derive a Wasserstein distance between density matrices and matrix-valued distributions.

Similar remarks apply to the vector-valued case in which one must also invoke some ideas from graph theory in formulating our generalization of scalar-valued densities. See [12] for all the details. For another approach to extending the Benamou-Brenier theory to graphs, see [16]. The extension of mass transport theory to vector-valued data may also have far-reaching consequences for several key problems, e.g., combining genomic and proteomic networks in medical applications [35].

In this paper, we focus on algorithms for the numerical solution of the optimal matrix-valued mass transport problems introduced in [9, 11, 26], and the vector-valued case formulated in [12]. In [11, 12], both problems are reformulated as convex optimization problems. We adopt an inexact sequential quadratic programming (SQP) method [8, 29, 34] to tackle such convex optimization problems. Similar methods have been applied to scalar optimal mass transport [18]. Recently, an efficient algorithm for the L^1 counterpart of vector-valued and matrix-valued mass transport problems was studied in [33].

The remainder of this paper is summarized as follows. Section 2 is a brief introduction to the matrix-valued optimal transport theory. We develop the corresponding algorithm in Sect. 3, and then the algorithm for vector-valued optimal transport is described in Sect. 4. We conclude with several illustrative examples to demonstrate our algorithm in Sect. 5.

2 Matrix-Valued Optimal Mass Transport

In this section, we sketch the approach [11] for which the convex optimization algorithm given in the present note was formulated. As noted above, similar approaches to matrix-valued OMT were formulated independently in [9, 26].

2.1 Gradient on Space of Hermitian Matrices

Denote by $\mathcal{H}$ and $\mathcal{S}$ the set of $n \times n$ Hermitian and skew-Hermitian matrices, respectively. We will assume that all of our matrices are of fixed size $n \times n$. Next, we denote the space of block-column vectors consisting of N elements in $\mathcal{S}$ and $\mathcal{H}$ as $\mathcal{S}^N$ and $\mathcal{H}^N$, respectively. We also let $\mathcal{H}_+$ and $\mathcal{H}_{++}$ denote the cones of nonnegative and positive-definite matrices, respectively, and we use the standard notion of inner product, namely

$$\langle X, Y \rangle = \operatorname{tr}(X^* Y),$$

for both $\mathcal{H}$ and $\mathcal{S}$. For $X, Y \in \mathcal{H}^N(\mathcal{S}^N)$,

$$\langle X, Y \rangle = \sum_{k=1}^N \operatorname{tr}(X_k^* Y_k).$$

Given $X = [X_1^*, \dots, X_N^*]^* \in \mathcal{H}^N(\mathcal{S}^N)$, $Y \in \mathcal{H}(\mathcal{S})$, set

$$XY = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix} Y := \begin{bmatrix} X_1 Y \\ \vdots \\ X_N Y \end{bmatrix},$$

$$YX = Y \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix} := \begin{bmatrix} YX_1 \\ \vdots \\ YX_N \end{bmatrix},$$

and

$$\bar{X} = \begin{bmatrix} X_1^* \\ \vdots \\ X_N^* \end{bmatrix}.$$

For a given $L \in \mathcal{H}^N$ we define

$$\nabla_L : \mathcal{H} \rightarrow \mathcal{S}^N, \quad X \mapsto \begin{bmatrix} L_1 X - X L_1 \\ \vdots \\ L_N X - X L_N \end{bmatrix} \quad (1)$$

to be the **gradient operator** [9, 12]. By analogy with the ordinary multivariable calculus, we refer to its dual with respect to the Hilbert-Schmidt inner product as the (negative) **divergence operator**, and this is

$$\nabla_L^* : \mathcal{S}^N \rightarrow \mathcal{H}, \quad Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix} \mapsto \sum_k^N L_k Y_k - Y_k L_k, \quad (2)$$

i.e., ∇_L^* is defined by means of the identity

$$\langle \nabla_L X, Y \rangle = \langle X, \nabla_L^* Y \rangle.$$

A standing assumption throughout, is that the null space of ∇_L , denoted by $\ker(\nabla_L)$, contains only scalar multiples of the identity matrix. In this note, we use one such basis generated by the following $N = 2$ components:

$$L_1 = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \end{bmatrix}, \quad L_2 = \text{diag}([1, 2, \dots, n-1, 0]).$$

It should be noted that a different basis L will lead to a different metric and a different result.

2.2 Matrix-Valued Optimal Mass Transport

We next sketch the formulation for matrix-valued optimal mass transport proposed in [11]. Given a convex compact set $E \subset \mathbb{R}^m$, denote

$$\mathcal{D} = \left\{ \rho(\cdot) \in \mathcal{H}_+ \mid \int_E \text{tr}(\rho(x)) dx = 1 \right\},$$

$$\mathcal{D}_+ = \left\{ \rho(\cdot) \in \mathcal{H}_{++} \mid \int_E \text{tr}(\rho(x)) dx = 1 \right\}.$$

Let $\rho^0, \rho^1 \in \mathcal{D}_+$ be two matrix-valued densities defined on E with positive values. A dynamic formulation of matrix-valued optimal mass transport between these two given marginals is [11],

$$\min_{\rho \in \mathcal{D}_+, w \in \mathcal{H}^m, v \in \mathcal{S}^N} \int_0^1 \int_E \{ \text{tr}(\rho w^* w) + \gamma \text{tr}(\rho v^* v) \} dx dt, \quad (3a)$$

$$\text{subject to } \frac{\partial \rho}{\partial t} + \frac{1}{2} \nabla_x \cdot (w\rho + \rho w) - \frac{1}{2} \nabla_L^* (v\rho + \rho v) = 0, \quad (3b)$$

$$\rho(0, \cdot) = \rho^0, \quad \rho(1, \cdot) = \rho^1 \quad (3c)$$

with $\nabla_x \cdot$ being the standard divergence operator in $\mathbb{R}^m$. We impose a Neumann zero-flux boundary condition on the flux $w\rho$. Note no boundary condition is needed for $v\rho$ since ∇_L^* is applied pointwise. By defining $p = w\rho, u = v\rho$, the above can be cast as a convex optimization problem

$$\min_{\rho, p, u} \int_0^1 \int_E \{ \text{tr}(p\rho^{-1}p^*) + \gamma \text{tr}(u\rho^{-1}u^*) \} dx dt, \quad (4a)$$

$$\text{subject to } \frac{\partial \rho}{\partial t} + \frac{1}{2} \nabla_x \cdot (p + \bar{p}) - \frac{1}{2} \nabla_L^* (u - \bar{u}) = 0, \quad (4b)$$

$$\rho(0, \cdot) = \rho^0, \quad \rho(1, \cdot) = \rho^1. \quad (4c)$$

We remark that $(p + \bar{p})/2 \in \mathcal{H}^m$ and $(u - \bar{u})/2 \in \mathcal{S}^N$, which is consistent with the domain of ∇_L^* . For the sake of brevity, the set E is taken to be the unit cube $[0, 1]^m$.

3 Discretization and Algorithm: Matrix-Valued Case

We follow closely the algorithm developed in [18] for scalar optimal mass transport problems. We restrict ourselves to the real-valued case, that is, $\mathcal{H}$ and $\mathcal{S}$ denote symmetric and skew-symmetric matrices, respectively. In order to highlight the key parts of our methodology, we first consider the discretization for the 1D case, i.e., $m = 1$. In particular, we take $E = [0, 1]$. The algorithm extends almost verbatim to the higher dimensional setting as we will see in Sect. 3.4.

We discretize the space-time domain $[0, 1] \times [0, 1]$ into $n_x \times n_t$ rectangular cells. Denote Ω_{ij} , $1 \leq i \leq n_x$, $1 \leq j \leq n_t$ as the (i, j) box. As shown in Fig. 1, we use a staggered grid to discretize p and ρ . **The variable u is, however, valued at the centers of the cells $\{\Omega_{ij}\}$.** This is due to the fact the quantum gradient operator ∇_L doesn't involve time or space. More specifically,

$$p = \left(p_{i+\frac{1}{2},j} \right), \quad 0 \leq i \leq n_x, \quad 1 \leq j \leq n_t$$

$$\rho = \left(\rho_{i,j+\frac{1}{2}} \right), \quad 1 \leq i \leq n_x, \quad 0 \leq j \leq n_t$$

$$u = (u_{i,j}), \quad 1 \leq i \leq n_x, \quad 1 \leq j \leq n_t.$$

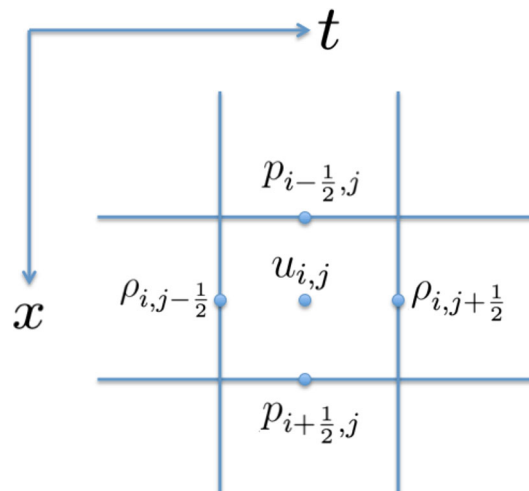
Note that $p_{i+\frac{1}{2},j} \in \mathbb{C}^{n \times n}$, $\rho_{i,j+\frac{1}{2}} \in \mathcal{H}_+$, $u_{i,j} \in \mathbb{C}^{n \times n}$. The boundary values are

$$p_{\frac{1}{2},j} = 0, \quad p_{n_x+\frac{1}{2},j} = 0, \quad 1 \leq j \leq n_t$$

and

$$\rho_{i,\frac{1}{2}} = \rho_i^0, \quad \rho_{i,n_t+\frac{1}{2}} = \rho_i^1, \quad 1 \leq i \leq n_x.$$

Fig. 1 Ω_{ij}



We exclude the boundary values from the variables and denote

$$p = \left(p_{i+\frac{1}{2},j} \right), \quad 1 \leq i \leq n_x - 1, \quad 1 \leq j \leq n_t$$

$$\rho = \left(\rho_{i,j+\frac{1}{2}} \right), \quad 1 \leq i \leq n_x, \quad 1 \leq j \leq n_t - 1.$$

3.1 Continuity Equation: Matrix-Valued Case

We use the above discretizing scheme, together with the boundary conditions to rewrite the continuity Eq. (4b) as

$$D_1 p + D_2 \rho + D_3 u = b. \quad (5)$$

Here the linear operators D_1, D_2, D_3 are defined as

$$(D_1 p)_{i,j} = \begin{cases} \frac{1}{2} \left(p_{i+\frac{1}{2},j} + p_{i+\frac{1}{2},j}^* - p_{i-\frac{1}{2},j} - p_{i-\frac{1}{2},j}^* \right) / h_x, & \text{if } 2 \leq i \leq n_x - 1, \\ \frac{1}{2} \left(p_{\frac{3}{2},j} + p_{\frac{3}{2},j}^* \right) / h_x, & \text{if } i = 1, \\ -\frac{1}{2} \left(p_{n_x-\frac{1}{2},j} + p_{n_x-\frac{1}{2},j}^* \right) / h_x, & \text{if } i = n_x, \end{cases}$$

$$(D_2 \rho)_{i,j} = \begin{cases} \left(\rho_{i,j+\frac{1}{2}} - \rho_{i,j-\frac{1}{2}} \right) / h_t, & \text{if } 2 \leq j \leq n_t - 1, \\ \rho_{i,\frac{3}{2}} / h_t, & \text{if } j = 1, \\ -\rho_{i,n_t-\frac{1}{2}} / h_t, & \text{if } j = n_t, \end{cases}$$

$$(D_3 u)_{i,j} = -\frac{1}{2} \nabla_L^* (u_{i,j} - \bar{u}_{i,j}), \quad 1 \leq i \leq n_x, \quad 1 \leq j \leq n_t.$$

The matrices $b = (b_{i,j}), 1 \leq i \leq n_x, 1 \leq j \leq n_t$ carries the information of the boundary values ρ^0 and ρ^1 . More specifically,

$$b_{i,j} = \begin{cases} \rho_i^0 / h_t & \text{if } j = 1, \\ -\rho_i^1 / h_t & \text{if } j = n_t, \\ 0 & \text{otherwise.} \end{cases}$$

3.2 Discretizing the Cost Function: Matrix-Valued Case

We use a combination of midpoint and trapezoidal methods to discretize the cost function. On the volume Ω_{ij} we have

$$\begin{aligned} \int_{\Omega_{ij}} \{ \text{tr} (p \rho^{-1} p^*) + \gamma \text{tr} (u \rho^{-1} u^*) \} &\approx \frac{h_x h_t}{4} \text{tr} \left(\left(p_{i-\frac{1}{2},j}^* p_{i-\frac{1}{2},j} + p_{i+\frac{1}{2},j}^* p_{i+\frac{1}{2},j} \right) \right. \\ &\quad \left. \left(\rho_{i,j-\frac{1}{2}}^{-1} + \rho_{i,j+\frac{1}{2}}^{-1} \right) \right) \\ &\quad + \frac{\gamma h_x h_t}{2} \text{tr} \left(u_{i,j}^* u_{i,j} \left(\rho_{i,j-\frac{1}{2}}^{-1} + \rho_{i,j+\frac{1}{2}}^{-1} \right) \right). \end{aligned}$$

Let A_1 be the averaging operator over the spatial domain and A_2 be the averaging operator over the time domain (one needs to be careful about the boundaries). Then the cost function (4a) may be approximated by

$$\langle A_1(p^* \circ p), A_2(\rho^{-1}) + a \rangle h_x h_t + \langle u^* \circ u, A_2(\rho^{-1}) + a \rangle \gamma h_x h_t, \quad (6)$$

where $a \geq 0$ depends only on the boundary values ρ^0 and ρ^1 . The inverse operator and the multiplication operator $\circ$ are applied block-wise. More specifically, the expressions for A_1, A_2, a are

$$\begin{aligned} (A_1(p^* \circ p))_{i,j} &= \begin{cases} \frac{1}{2} \left(p_{i-\frac{1}{2},j}^* p_{i-\frac{1}{2},j} + p_{i+\frac{1}{2},j}^* p_{i+\frac{1}{2},j} \right), & \text{if } 2 \leq i \leq n_x - 1, \\ \frac{1}{2} p_{\frac{3}{2},j}^* p_{\frac{3}{2},j}, & \text{if } i = 1, \\ \frac{1}{2} p_{n_x-\frac{1}{2},j}^* p_{n_x-\frac{1}{2},j}, & \text{if } i = n_x, \end{cases} \\ (A_2(\rho^{-1}))_{i,j} &= \begin{cases} \frac{1}{2} \left(\rho_{i,j-\frac{1}{2}}^{-1} + \rho_{i,j+\frac{1}{2}}^{-1} \right), & \text{if } 2 \leq j \leq n_t - 1, \\ \frac{1}{2} \rho_{i,\frac{3}{2}}^{-1}, & \text{if } j = 1, \\ \frac{1}{2} \rho_{i,n_t-\frac{1}{2}}^{-1}, & \text{if } j = n_t, \end{cases} \\ a_{i,j} &= \begin{cases} \frac{1}{2} (\rho_i^0)^{-1} & \text{if } j = 1, \\ \frac{1}{2} (\rho_i^1)^{-1} & \text{if } j = n_t, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

We remark that it is important to first square then average, and first invert then average, to guarantee stability [3, 18].

3.3 Sequential Quadratic Programming (SQP)

Following the above discretization scheme, we obtain the discrete convex optimization problem

$$\min \quad f(p, \rho, u) = \langle A_1(p^* p), A_2(\rho^{-1}) + a \rangle h_x h_t + \langle u^* u, A_2(\rho^{-1}) + a \rangle \gamma h_x h_t, \quad (7a)$$

$$\text{subject to} \quad D_1 p + D_2 \rho + D_3 u = b. \quad (7b)$$

The Lagrangian of this problem is

$$\mathcal{L}(p, \rho, u) = f(p, \rho, u) / (h_x h_t) + \langle \lambda, D_1 p + D_2 \rho + D_3 u - b \rangle.$$

The KKT condition [8, 29]

$$\nabla_p \mathcal{L} = D_1^* \lambda + 2p \circ A_1^* (A_2(\rho^{-1}) + a) = 0 \quad (8a)$$

$$\nabla_\rho \mathcal{L} = D_2^* \lambda - \rho^{-1} \circ A_2^* A_1(p^* p) \circ \rho^{-1} - \gamma \rho^{-1} \circ A_2^* (u^* u) \circ \rho^{-1} = 0 \quad (8b)$$

$$\nabla_u \mathcal{L} = D_3^* \lambda + 2\gamma u \circ (A_2(\rho^{-1}) + a) = 0 \quad (8c)$$

$$\nabla_\lambda \mathcal{L} = D_1 p + D_2 \rho + D_3 u - b = 0 \quad (8d)$$

follow, with $\circ$ denoting block-wise multiplication.

Let $w = (p, \rho, u)$, $D = (D_1, D_2, D_3)$, then at each SQP iteration we solve the system

$$\begin{pmatrix} \hat{A} & D^* \\ D & 0 \end{pmatrix} \begin{pmatrix} \delta w \\ \delta \lambda \end{pmatrix} = - \begin{pmatrix} \nabla_w \mathcal{L} \\ \nabla_\lambda \mathcal{L} \end{pmatrix}, \quad (9)$$

and update w, λ using line search. In (9), by slight abuse of notation, $w, \lambda, \nabla_w \mathcal{L}, \nabla_\lambda \mathcal{L}$ are vectorizations of the corresponding matrices and D is the matrix version of the linear operator

D . In principle, Problem 7 can be solved using Newton's method. However, the mixed terms introduce off-diagonal elements in the Hessian, which makes it forbidden for large problems. We adopt an inexact SQP method [8]. The matrix $\hat{A}$ is an approximation of the Hessian of the objective function

$$\hat{A} = \begin{pmatrix} 2\text{Bdiag}(A_1^*(A_2(\rho^{-1}) + a)) & 0 & 0 \\ 0 & \text{Bdiag}(g(p, \rho, u)) & 0 \\ 0 & 0 & 2\gamma\text{Bdiag}(A_2(\rho^{-1}) + a) \end{pmatrix}.$$

Here Bdiag denotes block diagonal operator. More specifically,

$$\text{Bdiag}([T_1, T_2, \dots, T_k]) = \begin{bmatrix} T_1 & 0 & \cdots & 0 \\ 0 & T_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & T_k \end{bmatrix}$$

for a block vector of linear operators $[T_1, T_2, \dots, T_k]$. The operator $g(p, \rho, u)$ is the Hessian of f over ρ with $g_{i,j+\frac{1}{2}}$ being the map

$$\begin{aligned} g_{i,j+\frac{1}{2}}(X) = & \rho_{i,j+\frac{1}{2}}^{-1} (A_2^* A_1 (p^* p))_{i,j+\frac{1}{2}} \rho_{i,j+\frac{1}{2}}^{-1} X \rho_{i,j+\frac{1}{2}}^{-1} \\ & + \rho_{i,j+\frac{1}{2}}^{-1} X \rho_{i,j+\frac{1}{2}}^{-1} (A_2^* A_1 (p^* p))_{i,j+\frac{1}{2}} \rho_{i,j+\frac{1}{2}}^{-1} \\ & + \gamma \rho_{i,j+\frac{1}{2}}^{-1} (A_2^* (u^* u))_{i,j+\frac{1}{2}} \rho_{i,j+\frac{1}{2}}^{-1} X \rho_{i,j+\frac{1}{2}}^{-1} \\ & + \gamma \rho_{i,j+\frac{1}{2}}^{-1} X \rho_{i,j+\frac{1}{2}}^{-1} (A_2^* (u^* u))_{i,j+\frac{1}{2}} \rho_{i,j+\frac{1}{2}}^{-1}. \end{aligned}$$

In each step we solve the linear system (9) in an inexact manner. There are many methods to achieve this. In our approach, we apply the Schur complement and solve the reduced system

$$D\hat{A}^{-1}D^*\delta\lambda = \nabla_\lambda \mathcal{L} - D\hat{A}^{-1}\nabla_w \mathcal{L}$$

using preconditioned conjugated gradients method with incomplete Cholesky factorization [21] as a preconditioner. The matrix D is essentially a discretization of a divergence operator. Therefore, solving the above system is equivalent to solving a Poisson equation. The update for w is then given by

$$\delta w = -\hat{A}^{-1} (D^*\delta\lambda + \nabla_w \mathcal{L}).$$

A merit function based line search is used [29]. Moreover, the step size is chosen to be sufficiently small so that the positivity of ρ is preserved. One alternative approach to guarantee the positivity of ρ is via a projection method as discussed in [32]. One may also add regularization terms as described in [14, 23].

Remark 1 In our numerical implementation, we take advantage of the structure of ρ being symmetric, and only save the upper triangular part of it. This is beneficial in terms of both memory and speed.

3.4 Higher Dimensions: Matrix-Valued Case

In this section we work out higher dimensional cases, namely 2D and 3D. In practice, for the applications we have in mind, 3D suffices.

We begin with the 2D case. Accordingly, we have the discrete convex optimization problem

$$\begin{aligned} \min \quad & f(p, \rho, u) = \left\langle A_{1x}(p_x^* p_x) + A_{1y}(p_y^* p_y), A_2(\rho^{-1}) + a \right\rangle h_x h_y h_t \\ & + \langle u^* u, A_2(\rho^{-1}) + a \rangle \gamma h_x h_y h_t \\ \text{subject to} \quad & D_{1x} p_x + D_{1y} p_y + D_2 \rho + D_3 u = b. \end{aligned}$$

The Lagrangian of this problem is

$$\mathcal{L}(p, \rho, u) = f(p, \rho, u) / (h_x h_y h_t) + \langle \lambda, D_{1x} p_x + D_{1y} p_y + D_2 \rho + D_3 u - b \rangle.$$

In the above,

$$a_{i,j,k} = \begin{cases} \frac{1}{2} (\rho_{i,j}^0)^{-1} & \text{if } k = 1, \\ \frac{1}{2} (\rho_{i,j}^1)^{-1} & \text{if } k = n_t, \\ 0 & \text{otherwise.} \end{cases}$$

and

$$b_{i,j,k} = \begin{cases} \rho_{i,j}^0 / h_t & \text{if } k = 1, \\ -\rho_{i,j}^1 / h_t & \text{if } k = n_t, \\ 0 & \text{otherwise.} \end{cases}$$

It follows that the KKT conditions are

$$\nabla_{p_x} \mathcal{L} = D_{1x}^* \lambda + 2p_x \circ A_{1x}^* (A_2(\rho^{-1}) + a) = 0 \quad (10a)$$

$$\nabla_{p_y} \mathcal{L} = D_{1y}^* \lambda + 2p_y \circ A_{1y}^* (A_2(\rho^{-1}) + a) = 0 \quad (10b)$$

$$\begin{aligned} \nabla_{\rho} \mathcal{L} &= D_2^* \lambda - \rho^{-1} \circ A_2^* (A_{1x}(p_x^* p_x) + A_{1y}(p_y^* p_y)) \circ \rho^{-1} \\ &\quad - \gamma \rho^{-1} \circ A_2^* (u^* u) \circ \rho^{-1} = 0 \end{aligned} \quad (10c)$$

$$\nabla_u \mathcal{L} = D_3^* \lambda + 2\gamma u \circ (A_2(\rho^{-1}) + a) = 0 \quad (10d)$$

$$\nabla_{\lambda} \mathcal{L} = D_1 p + D_2 \rho + D_3 u - b = 0, \quad (10e)$$

with $\circ$ denoting block-wise multiplication as before.

Let $w = (p_x, p_y, \rho, u)$. Then at each SQP iteration, we solve the system

$$\begin{pmatrix} \hat{A} & D^* \\ D & 0 \end{pmatrix} \begin{pmatrix} \delta w \\ \delta \lambda \end{pmatrix} = - \begin{pmatrix} \nabla_w \mathcal{L} \\ \nabla_{\lambda} \mathcal{L} \end{pmatrix}, \quad (11)$$

where $D = (D_{1x}, D_{1y}, D_2, D_3)$. The matrix $\hat{A}$ is an approximation of the Hessian of the objective function

$$\hat{A} = \begin{pmatrix} 2\text{Bdiag}(A_{1x}^* (A_2(\rho^{-1}) + a)) & 0 & 0 & 0 \\ 0 & 2\text{Bdiag}(A_{1y}^* (A_2(\rho^{-1}) + a)) & 0 & 0 \\ 0 & 0 & \text{Bdiag}(g(p, \rho, u)) & 0 \\ 0 & 0 & 0 & 2\gamma \text{Bdiag}(A_2(\rho^{-1}) + a) \end{pmatrix}$$

The operator $g(p, \rho, u)$ is the Hessian of f over ρ with $g_{i,j,k+\frac{1}{2}}$ being the map

$$g_{i,j,k+\frac{1}{2}}(X) = \rho_{i,j,k+\frac{1}{2}}^{-1} \left(A_2^* \left(A_{1x} (p_x^* p_x) + A_{1y} (p_y^* p_y) + \gamma u^* u \right) \right)_{i,j,k+\frac{1}{2}} \rho_{i,j,k+\frac{1}{2}}^{-1} X \rho_{i,j,k+\frac{1}{2}}^{-1} \\ + \rho_{i,j,k+\frac{1}{2}}^{-1} X \rho_{i,j,k+\frac{1}{2}}^{-1} \left(A_2^* \left(A_{1x} (p_x^* p_x) + A_{1y} (p_y^* p_y) + \gamma u^* u \right) \right)_{i,j,k+\frac{1}{2}} \rho_{i,j,k+\frac{1}{2}}^{-1}.$$

The 3D case is quite similar. Now, we have the discrete convex optimization problem

$$\begin{aligned} \min \quad & f(p, \rho, u) = \left\langle A_{1x} (p_x^* p_x) + A_{1y} (p_y^* p_y) + A_{1z} (p_z^* p_z), A_2(\rho^{-1}) + a \right\rangle h_x h_y h_z h_t \\ & + \left\langle u^* u, A_2(\rho^{-1}) + a \right\rangle \gamma h_x h_y h_z h_t \\ \text{subject to} \quad & D_{1x} p_x + D_{1y} p_y + D_{1z} p_z + D_2 \rho + D_3 u = b. \end{aligned}$$

The Lagrangian of this problem is

$$\mathcal{L}(p, \rho, u) = f(p, \rho, u) / (h_x h_y h_z h_t) + \langle \lambda, D_{1x} p_x + D_{1y} p_y + D_{1z} p_z + D_2 \rho + D_3 u - b \rangle.$$

In the above,

$$a_{i,j,k,\ell} = \begin{cases} \frac{1}{2} (\rho_{i,j,k}^0)^{-1} & \text{if } \ell = 1, \\ \frac{1}{2} (\rho_{i,j,k}^1)^{-1} & \text{if } \ell = n_t, \\ 0 & \text{otherwise.} \end{cases}$$

and

$$b_{i,j,k,\ell} = \begin{cases} \rho_{i,j,k}^0 / h_t & \text{if } \ell = 1, \\ -\rho_{i,j,k}^1 / h_t & \text{if } \ell = n_t, \\ 0 & \text{otherwise.} \end{cases}$$

It follows that the KKT conditions now are

$$\begin{aligned} \nabla_{p_x} \mathcal{L} &= D_{1x}^* \lambda + 2 p_x \circ A_{1x}^* (A_2(\rho^{-1}) + a) = 0 \\ \nabla_{p_y} \mathcal{L} &= D_{1y}^* \lambda + 2 p_y \circ A_{1y}^* (A_2(\rho^{-1}) + a) = 0 \\ \nabla_{p_z} \mathcal{L} &= D_{1z}^* \lambda + 2 p_z \circ A_{1z}^* (A_2(\rho^{-1}) + a) = 0 \\ \nabla_{\rho} \mathcal{L} &= D_2^* \lambda - \rho^{-1} \circ A_2^* (A_{1x} (p_x^* p_x) + A_{1y} (p_y^* p_y) + A_{1z} (p_z^* p_z)) \circ \rho^{-1} \\ &\quad - \gamma \rho^{-1} \circ A_2^* (u^* u) \circ \rho^{-1} = 0 \\ \nabla_u \mathcal{L} &= D_3^* \lambda + 2 \gamma u \circ (A_2(\rho^{-1}) + a) = 0 \\ \nabla_{\lambda} \mathcal{L} &= D_1 p + D_2 \rho + D_3 u - b = 0, \end{aligned}$$

with $\circ$ the block-wise multiplication as earlier.

Let $w = (p_x, p_y, p_z, \rho, u)$, then at each SQP iteration we solve the system

$$\begin{pmatrix} \hat{A} & D^* \\ D & 0 \end{pmatrix} \begin{pmatrix} \delta w \\ \delta \lambda \end{pmatrix} = - \begin{pmatrix} \nabla_w \mathcal{L} \\ \nabla_{\lambda} \mathcal{L} \end{pmatrix}, \quad (12)$$

where $D = (D_{1x}, D_{1y}, D_{1z}, D_2, D_3)$. The matrix $\hat{A}$ is an approximation of the Hessian of the objective function

$$\begin{pmatrix} 2\text{Bdiag}(A_{1x}^*(A_2(\rho^{-1}) + a)) & 0 & 0 \\ 0 & 2\text{Bdiag}(A_{1y}^*(A_2(\rho^{-1}) + a)) & 0 \\ 0 & 0 & 2\text{Bdiag}(A_{1z}^*(A_2(\rho^{-1}) + a)) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \text{Bdiag}(g(p, \rho, u)) & 0 & 0 \\ 0 & 2\gamma\text{Bdiag}(A_2(\rho^{-1}) + a) & 0 \end{pmatrix}$$

The operator $g(p, \rho, u)$ is the Hessian of f over ρ with $g_{i,j,k,\ell+\frac{1}{2}}$ being the map

$$\begin{aligned} g_{i,j,k,\ell+\frac{1}{2}}(X) = & \rho_{i,j,k,\ell+\frac{1}{2}}^{-1} \left(A_2^* \left(A_{1x}(p_x^* p_x) + A_{1y}(p_y^* p_y) + A_{1z}(p_z^* p_z) \right. \right. \\ & \left. \left. + \gamma u^* u \right) \right)_{i,j,k,\ell+\frac{1}{2}} \rho_{i,j,k,\ell+\frac{1}{2}}^{-1} X \rho_{i,j,k,\ell+\frac{1}{2}}^{-1} \\ & + \rho_{i,j,k,\ell+\frac{1}{2}}^{-1} X \rho_{i,j,k,\ell+\frac{1}{2}}^{-1} \left(A_2^* \left(A_{1x}(p_x^* p_x) + A_{1y}(p_y^* p_y) + A_{1z}(p_z^* p_z) \right. \right. \\ & \left. \left. + \gamma u^* u \right) \right)_{i,j,k,\ell+\frac{1}{2}} \rho_{i,j,k,\ell+\frac{1}{2}}^{-1} \end{aligned}$$

4 Vector-Valued Optimal Mass Transport

Next we move to vector-valued optimal transport, which was proposed recently in [12]. We briefly review the setup in this section, and refer the reader to [12] for details.

4.1 Gradients on Graphs

We consider a connected, positively weighted, undirected graph $\mathcal{F} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ with n nodes labeled as i , with $1 \leq i \leq n$, and N edges. We have that $\Delta_{\mathcal{F}} = -\mathbb{D}W\mathbb{D}^T$ where $\Delta_{\mathcal{F}}, \mathbb{D}, W = \text{diag}\{\mathbf{w}_1, \dots, \mathbf{w}_N\}$ are the graph Laplacian, incidence, and weight matrices, respectively. One can define the Laplacian in terms of a graph gradient and divergence as

$$\Delta_{\mathcal{F}} = -\nabla_{\mathcal{F}}^* \nabla_{\mathcal{F}},$$

where

$$\nabla_{\mathcal{F}} : \mathbb{R}^n \rightarrow \mathbb{R}^N, \quad x \mapsto W^{1/2} \mathbb{D}^T x$$

denotes the gradient operator and

$$\nabla_{\mathcal{F}}^* : \mathbb{R}^N \rightarrow \mathbb{R}^n, \quad y \mapsto \mathbb{D}W^{1/2} y$$

denotes its dual.

4.2 Vector-Valued Optimal Mass Transport

We begin by considering a *vector-valued density* ρ on $\mathbb{R}^m$, i.e., a map from $E \subset \mathbb{R}^m$ to $\mathbb{R}_+^n$ such that

$$\sum_{i=1}^n \int_E \rho_i(x) dx = 1.$$

Here the convex compact set $E \subset \mathbb{R}^m$ is a domain where the densities are defined, typically the unit n -dimensional cube. To avoid proliferation of symbols, we denote the set of all vector-valued densities and its interior again by $\mathcal{D}$ and $\mathcal{D}_+$, respectively. We refer to the entries of ρ as representing density or mass of individual species/particles that can mutate between one another while maintaining total mass. Mass transfer may only be permissible between specific types of particles. Thus, allowable transfer can be modeled by the existence of a corresponding edge in a graph $\mathcal{F} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ whose vertices in $\mathcal{V}$ correspond to those individual species, see [12]. The edge weights in $\mathcal{W}$ can quantify cost, rate, or likelihood of transfer.

Following the arguments in [12], this leads to the following (symmetric) Wasserstein 2-metric on vector-valued distributions: Given two given marginals $\rho^0, \rho^1 \in \mathcal{D}_+$ the (square) of the Wasserstein distance is given by:

$$\min_{\rho, p, u} \int_0^1 \int_E \left\{ p^T \operatorname{diag}(\rho)^{-1} p + \gamma u^T [\operatorname{diag}(\mathbb{D}_2^T \rho)^{-1} + \operatorname{diag}(\mathbb{D}_1^T \rho)^{-1}] u \right\} dx dt \quad (13a)$$

$$\text{subject to } \frac{\partial \rho}{\partial t} + \nabla_x \cdot p - \nabla_{\mathcal{F}}^* u = 0, \quad (13b)$$

$$\rho(0, \cdot) = \rho^0, \quad \rho(1, \cdot) = \rho^1. \quad (13c)$$

Here u is the “flux” on graphs, $p = [p_1, \dots, p_n]^T$ is the “momentum” (mass times velocity vector field), the matrix $\mathbb{D}_1$ is the portion of the incidence matrix $\mathbb{D}$ containing 1’s (sources), and $\mathbb{D}_2 = \mathbb{D}_1 - \mathbb{D}$ (sinks).

We note that the algorithm is very similar to that of the matrix case. For completeness, the numerical implementation is described in detail in the “Appendix” below.

5 Numerical Experiments

Several examples are provided in this section to illustrate the effectiveness of our algorithms. For matrix-valued densities, we present examples in both 2D and 3D settings. In contrast, only 2D examples are studied for vector-valued densities.

5.1 Matrix-Valued Case

One motivation for matrix-valued optimal mass transport comes from diffusion tensor imaging (DTI). This is a widely used technique in magnetic resonance imaging. In diffusion images, the information at each pixel is captured in an ellipsoid, i.e., a 3×3 positive definite matrix, in lieu of a nonnegative number. The ellipsoids describe useful information such as the orientations of the brain fibers.

We tested our algorithm on a synthetic data set with $n = 3$. The initial density is a disk positioned at the center of the square domain and all the ellipsoids are isotropic. The terminal density contains four quarter discs located at the corners of the square domain, and the four

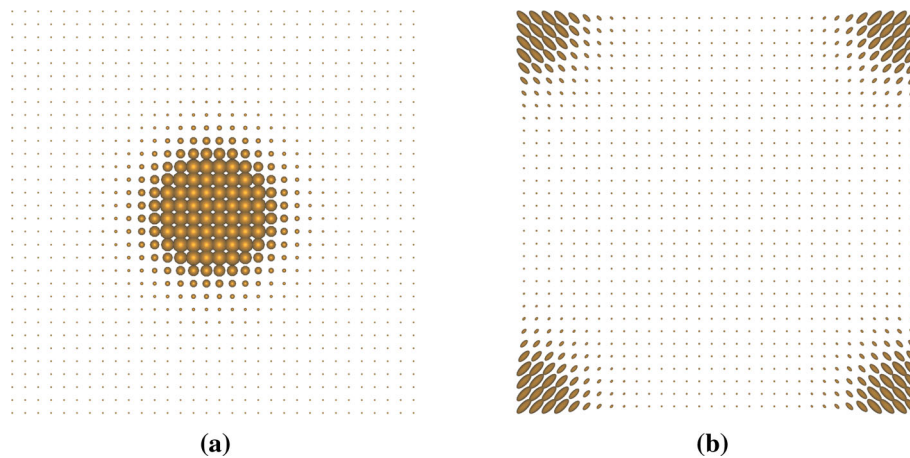


Fig. 2 Marginal distributions

components have different dominant directions. Both of them are depicted in Fig. 2. The densities have been smoothed to have low density contrast 10. Here the density contrast is defined to be the maximum of the ratios between the eigenvalues at different locations. In Fig. 3, we show the optimal density flow with grid size $32 \times 32 \times 10$ in space-time and parameter $\gamma = 0.01$. The masses split into four components and the ellipsoids change gradually from isotropic to anisotropic. However, the total mass is preserved along the entire interpolation as expected.

To demonstrate the performance of our algorithm, we tested it on the same problem with different mesh grid sizes: $16 \times 16 \times 10$, $32 \times 32 \times 20$, $64 \times 64 \times 40$ in space-time. We set the tolerance of the outer SQP iterations to 10^{-3} , and that of the preconditioning conjugate gradient solver in each iteration to 10^{-3} . The numbers of SQP iterations for convergence are shown in Table 1 for different mesh sizes.

We then studied the influence of density contrast and the parameter γ on the number of iterations needed to converge. The results for density contrast 50 are shown in Table 2 with tolerance 10^{-2} . We can see that the number of iterations increases as we increase the density contrast. Table 3 showcases the results for different γ values with fixed grid size $32 \times 32 \times 20$. We observe that the number of iterations is positively correlated with the value of γ .

Finally, we test our algorithm on a 3D data set. Table 4 displays the number of iterations for different grid sizes with density contrast 30 and parameter $\gamma = 0.1$.

5.2 Vector-Valued Case

An important application of vector-valued optimal mass transport is color image processing. In this cases, the vector-valued densities have three components corresponding to the intensities of the three basic colors red (R), green (G) and blue (B). The masses can transfer from one color channel to another and the cost of transferring is captured using a weighted graph $\mathcal{F}$. Here, we treat the three colors equally and take the graph to be a complete graph with unit weights, namely, $W = I$ and

$$\mathbb{D} = \begin{bmatrix} 1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & -1 \end{bmatrix}.$$

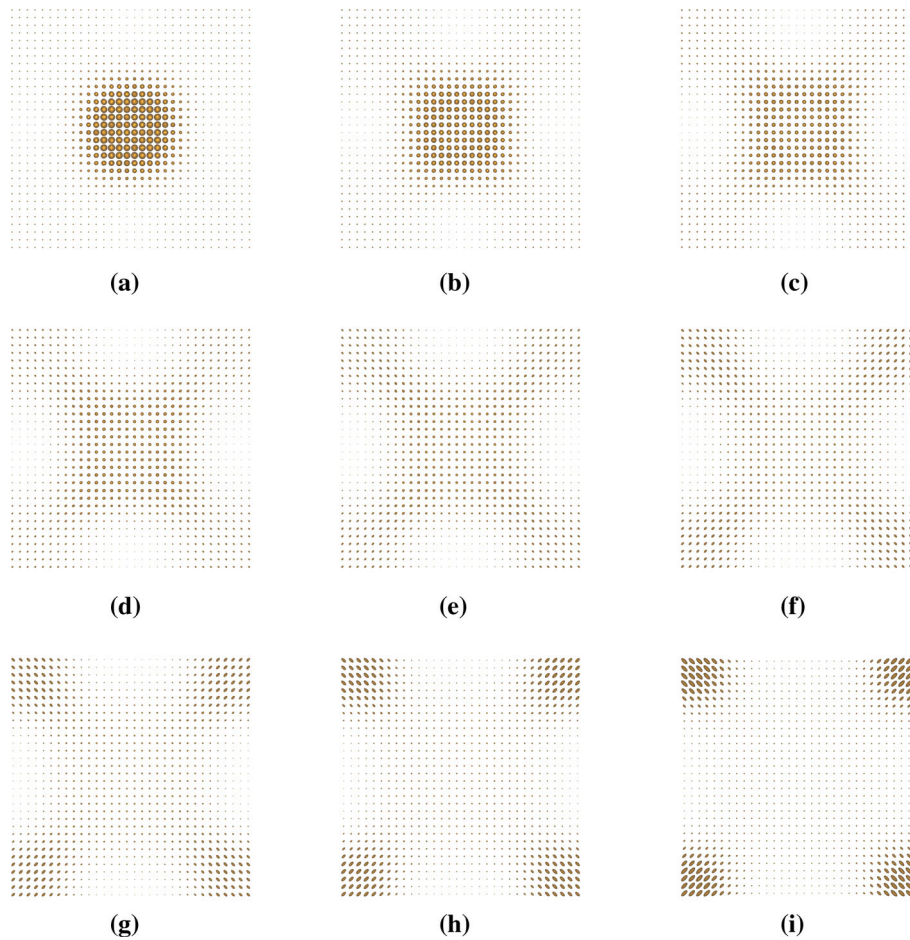


Fig. 3 Interpolation with $\gamma = 0.01$

The matrices $\mathbb{D}_1, \mathbb{D}_2$ in (13) are then

$$\mathbb{D}_1 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbb{D}_2 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}.$$

The two marginal densities are depicted in Fig. 4. The initial image ρ^0 is a disk located in the center of the square in white color, i.e., all three colors have equal intensity. The terminal distribution ρ^1 is an image of four circle quarters; one at each corner in different colors. Both the images have been smoothed to have density contrast $\max_k \sup_{x,y} \rho_k^i(x)/\rho_k^i(y) \approx 10$. Figure 5 illustrates the optimal interpolation using vector-valued optimal transport with grid size $128 \times 128 \times 10$ in space-time and parameter $\gamma = 0.01$. We observe that the white disk split into four circle quarters and meanwhile the colors change gradually from white to four different colors.

We next tested the performance of the algorithm with respect to the grid size. For this, we consider a grid hierarchy from a coarse grid of $32 \times 32 \times 10$ in space and time through a grid

Table 1 Number of SQP iterations required on different grid sizes for density contrast 10

Grid size	SQP iterations
$16 \times 16 \times 10$	19
$32 \times 32 \times 20$	27
$64 \times 64 \times 40$	35

Table 2 Number of SQP iterations required on different grid sizes for density contrast 50

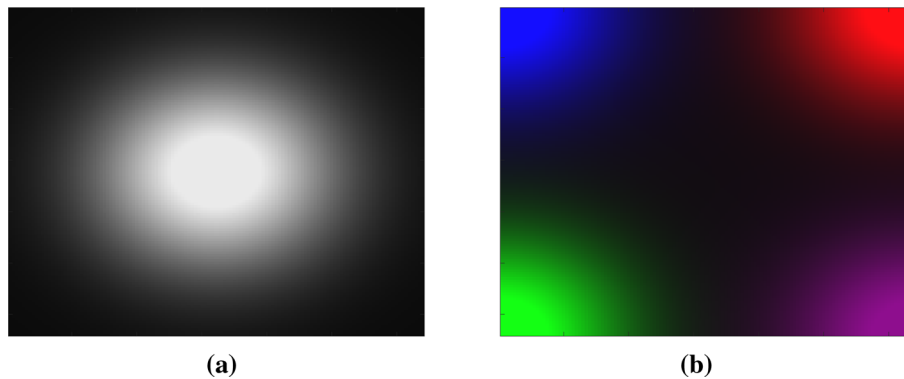
Grid size	SQP iterations
$16 \times 16 \times 10$	25
$32 \times 32 \times 20$	31
$64 \times 64 \times 40$	62

Table 3 Number of SQP iterations required for different γ

Parameter γ	SQP iterations
1	77
0.1	52
0.01	31

Table 4 Number of SQP iterations required on different grid sizes for 3D densities

Grid size	SQP iterations
$16 \times 16 \times 16 \times 10$	19
$32 \times 32 \times 32 \times 10$	25
$64 \times 64 \times 64 \times 10$	23

**Fig. 4** Marginal distributions

of $64 \times 64 \times 20$ to a grid of $128 \times 128 \times 40$. The parameter γ is set to be 0.01. The tolerance for the outer SQP iteration is set to be 10^{-3} and in each iteration the linear equation is solved with a relative residual of 10^{-2} . The numbers of SQP iterations are recorded in Table 5, from which we observe that the number of iterations needed doesn't increase much as we increase the size of the mesh grids.

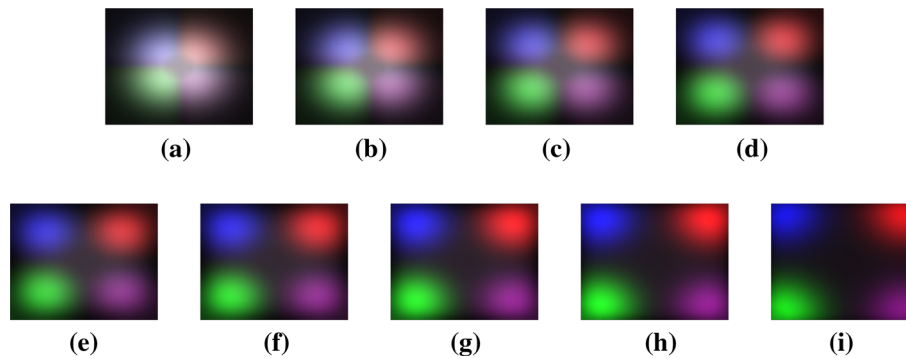


Fig. 5 Interpolation with $\gamma = 0.01$

Table 5 Number of SQP iterations required on different grid sizes for density contrast 10

Grid size	SQP iterations
$32 \times 32 \times 10$	11
$64 \times 64 \times 20$	12
$128 \times 128 \times 40$	14

Table 6 Number of SQP iterations required on different grid sizes for density contrast 100

Grid size	SQP iterations
$32 \times 32 \times 10$	24
$64 \times 64 \times 20$	27
$128 \times 128 \times 40$	32

Table 7 Number of SQP iterations required for different γ

Parameter γ	SQP iterations
1	48
0.1	42
0.01	27

We also applied the same algorithm to images with a higher density contrast 100. The results are shown in Table 6 for different grid sizes. As can be seen from the table, increasing the density contrast leads to an increasing of the number of SQP iterations. Again, the number of iterations needed to achieve certain precision is affected by the parameter. In Table 7 we display this change as a function of γ for fixed grid size $64 \times 64 \times 20$ and density contrast 100.

6 Conclusions and Future Work

In this paper, we described a fast algorithm for the numerical implementation of both matrix-valued and vector-valued versions of optimal mass transport. It is straightforward to extend this algorithm to cover matrix-valued transport problems with unequal masses (“unbalanced mass transport”) [13]. In the future, we intend to apply this methodology to various problems

including diffusion tensor magnetic resonance data, biological networks, and various types of vector-valued image data such as color and texture imagery. Finally, applying a multigrid methodology may speed up the linear solver even further, and will be a future direction in our research.

Acknowledgements This project was supported by AFOSR Grants (FA9550-15-1-0045 and FA9550-17-1-0435), ARO Grant (W911NF-17-1-049), grants from the National Center for Research Resources (P41-RR-013218) and the National Institute of Biomedical Imaging and Bioengineering (P41-EB-015902), National Science Foundation (NSF), grants from National Institutes of Health (1U24CA18092401A1, R01-AG048769), grants from the Swedish Research Council through the LCCC Linnaeus Center, and the Breast Cancer Research Foundation.

7 Appendix

As in the matrix-valued case, for simplicity of exposition, we consider the discretization in 1D case, and then describe the 2D case in Sect. 7.4 below. Thus, we take $E = [0, 1]$, and as before our technique extends almost verbatim to the higher dimensional setting. We should note that the algorithm presented here in the vector-valued case is very similar to the matrix optimal transport just described in the preceding sections. Nevertheless, for the sake of completeness, we present the key details here.

We discretize the space-time domain $[0, 1] \times [0, 1]$ into $n_x \times n_t$ rectangular cells. Denote Ω_{ij} , $1 \leq i \leq n_x$, $1 \leq j \leq n_t$ as the (i, j) box. We use staggered grid to discretize p and ρ . **The variable u is, however, valued at the centers of the cells $\{\Omega_{ij}\}$.** More specifically,

$$p = \left(p_{i+\frac{1}{2}, j} \right), \quad 0 \leq i \leq n_x, \quad 1 \leq j \leq n_t$$

$$\rho = \left(\rho_{i, j+\frac{1}{2}} \right), \quad 1 \leq i \leq n_x, \quad 0 \leq j \leq n_t$$

$$u = (u_{i, j}), \quad 1 \leq i \leq n_x, \quad 1 \leq j \leq n_t.$$

Note that the boundary values are

$$p_{\frac{1}{2}, j} = 0, \quad p_{n_x+\frac{1}{2}, j} = 0, \quad 1 \leq j \leq n_t$$

and

$$\rho_{i, \frac{1}{2}} = \rho_i^0, \quad \rho_{i, n_t+\frac{1}{2}} = \rho_i^1, \quad 1 \leq i \leq n_x.$$

We exclude the boundary values from the variables and denote

$$p = \left(p_{i+\frac{1}{2}, j} \right), \quad 1 \leq i \leq n_x - 1, \quad 1 \leq j \leq n_t$$

$$\rho = \left(\rho_{i, j+\frac{1}{2}} \right), \quad 1 \leq i \leq n_x, \quad 1 \leq j \leq n_t - 1.$$

7.1 Continuity Equation: Vector-Valued Case

We use the preceding discretizing scheme, together with the boundary conditions to rewrite the continuity Eq. (13b) as

$$D_1 p + D_2 \rho + D_3 u = b. \quad (14)$$

Here the linear operators D_1, D_2, D_3 are defined as

$$\begin{aligned} (D_1 p)_{i,j} &= \begin{cases} (p_{i+\frac{1}{2},j} - p_{i-\frac{1}{2},j}) / h_x, & \text{if } 2 \leq i \leq n_x - 1, \\ p_{\frac{3}{2},j} / h_x, & \text{if } i = 1, \\ -p_{n_x-\frac{1}{2},j} / h_x, & \text{if } i = n_x, \end{cases} \\ (D_2 \rho)_{i,j} &= \begin{cases} (\rho_{i,j+\frac{1}{2}} - \rho_{i,j-\frac{1}{2}}) / h_t, & \text{if } 2 \leq j \leq n_t - 1, \\ \rho_{i,\frac{3}{2}} / h_t, & \text{if } j = 1, \\ -\rho_{i,n_t-\frac{1}{2}} / h_t, & \text{if } j = n_t, \end{cases} \\ (D_3 u)_{i,j} &= -\nabla_{\mathcal{F}}^* u_{i,j}, \quad 1 \leq i \leq n_x, \quad 1 \leq j \leq n_t. \end{aligned}$$

The parameter b carries the information of the boundary values ρ^0 and ρ^1 . More specifically,

$$b_{i,j} = \begin{cases} \rho_i^0 / h_t & \text{if } j = 1, \\ -\rho_i^1 / h_t & \text{if } j = n_t, \\ 0 & \text{otherwise.} \end{cases}$$

7.2 Discretization of the Cost Function: Vector-Valued Case

Let A_1 be the averaging operator over the spatial domain and A_2 be the averaging operator over the time domain (as before one needs to be careful about the boundaries). Then the cost function (13a) may be approximated by

$$\langle A_1(p^2), A_2(1/\rho) + a \rangle h_x h_t + \langle u^2, A_2(1/(\mathbb{D}_2^T \rho) + 1/(\mathbb{D}_1^T \rho)) + c \rangle \gamma h_x h_t, \quad (15)$$

where $a \geq 0$ depends only on the boundary values ρ^0 and ρ^1 . The inverse operator and multiplication operators are applied block-wise. The expressions for A_1, A_2, a are

$$\begin{aligned} (A_1(p^2))_{i,j} &= \begin{cases} \frac{1}{2} (p_{i-\frac{1}{2},j}^2 + p_{i+\frac{1}{2},j}^2), & \text{if } 2 \leq i \leq n_x - 1, \\ \frac{1}{2} p_{\frac{3}{2},j}^2, & \text{if } i = 1, \\ \frac{1}{2} p_{n_x-\frac{1}{2},j}^2, & \text{if } i = n_x, \end{cases} \\ (A_2(1/\rho))_{i,j} &= \begin{cases} \frac{1}{2} (1/\rho_{i,j-\frac{1}{2}} + 1/\rho_{i,j+\frac{1}{2}}), & \text{if } 2 \leq j \leq n_t - 1, \\ 1/\rho_{i,\frac{3}{2}}, & \text{if } j = 1, \\ 1/\rho_{i,n_t-\frac{1}{2}}, & \text{if } j = n_t, \end{cases} \\ a_{i,j} &= \begin{cases} 1/\rho_i^0/2 & \text{if } j = 1, \\ 1/\rho_i^1/2 & \text{if } j = n_t, \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

$$c_{i,j} = \begin{cases} 1/\mathbb{D}_2^T \rho_i^0/2 + 1/\mathbb{D}_1^T \rho_i^0/2 & \text{if } j = 1, \\ 1/\mathbb{D}_2^T \rho_i^1/2 + 1/\mathbb{D}_1^T \rho_i^1/2 & \text{if } j = n_t, \\ 0 & \text{otherwise.} \end{cases}$$

7.3 Sequential Quadratic Programming (SQP)

From the above discussion, we obtain the discrete convex optimization problem

$$\min \quad f(p, \rho, u) = \langle A_1(p^2), A_2(1/\rho) + a \rangle h_x h_t + \langle u^2, A_2(1/(\mathbb{D}_2^T \rho) + 1/(\mathbb{D}_1^T \rho)) + c \rangle \gamma h_x h_t \quad (16a)$$

$$\text{subject to } D_1 p + D_2 \rho + D_3 u = b. \quad (16b)$$

The Lagrangian of this problem is

$$\mathcal{L}(p, \rho, u) = f(p, \rho, u) / (h_x h_t) + \langle \lambda, D_1 p + D_2 \rho + D_3 u - b \rangle.$$

It follows that the KKT conditions are given by

$$\nabla_p \mathcal{L} = D_1^T \lambda + 2p \circ A_1^T (A_2(1/\rho) + a) = 0 \quad (17a)$$

$$\begin{aligned} \nabla_\rho \mathcal{L} &= D_2^T \lambda - A_2^T A_1(p^2) / \rho^2 - \gamma \mathbb{D}_2 \left(A_2^T(u^2) / (\mathbb{D}_2^T \rho)^2 \right) \\ &\quad - \gamma \mathbb{D}_1 \left(A_2^T(u^2) / (\mathbb{D}_1^T \rho)^2 \right) = 0 \end{aligned} \quad (17b)$$

$$\nabla_u \mathcal{L} = D_3^T \lambda + 2\gamma u \circ \left(A_2(1/(\mathbb{D}_2^T \rho) + 1/(\mathbb{D}_1^T \rho)) + c \right) = 0 \quad (17c)$$

$$\nabla_\lambda \mathcal{L} = D_1 p + D_2 \rho + D_3 u - b = 0, \quad (17d)$$

with $\circ$ denoting block-wise multiplication.

Let $w = (p, \rho, u)$. Then at each SQP iteration, we solve the system

$$\begin{pmatrix} \hat{A} & D^T \\ D & 0 \end{pmatrix} \begin{pmatrix} \delta w \\ \delta \lambda \end{pmatrix} = - \begin{pmatrix} \nabla_w \mathcal{L} \\ \nabla_\lambda \mathcal{L} \end{pmatrix}, \quad (18)$$

where $D = (D_1, D_2, D_3)$. Again, the matrix $\hat{A}$ is an approximation of the Hessian of the objective function

$$\hat{A} = \begin{pmatrix} 2\text{diag}(A_1^T(A_2(1/\rho) + a)) & 0 & 0 \\ 0 & \text{diag}(g(p, \rho, u)) & 0 \\ 0 & 0 & 2\gamma \text{diag}(A_2(1/(\mathbb{D}_2^T \rho) + 1/(\mathbb{D}_1^T \rho)) + c) \end{pmatrix}.$$

The operator $g(p, \rho, u)$ is the Hessian of f over ρ with $g_{i,j+\frac{1}{2}}$ being the map

$$\begin{aligned} g_{i,j+\frac{1}{2}}(X) &= 2 \left(A_2^T A_1(p^2) \right)_{i,j+\frac{1}{2}} / \rho_{i,j+\frac{1}{2}}^3 X \\ &\quad + 2\gamma \mathbb{D}_2 \left[\left(A_2^T(u^2) \right)_{i,j+\frac{1}{2}} / (\mathbb{D}_2^T \rho)_{i,j+\frac{1}{2}}^3 \mathbb{D}_2^T X \right] \\ &\quad + 2\gamma \mathbb{D}_1 \left[\left(A_2^T(u^2) \right)_{i,j+\frac{1}{2}} / (\mathbb{D}_1^T \rho)_{i,j+\frac{1}{2}}^3 \mathbb{D}_1^T X \right]. \end{aligned}$$

7.4 Higher Dimensions: Vector-Valued Case

We concretely work out the 2D case in this section. The higher dimensional cases are very similar, but naturally involve additional indices. We should note that for all of the applications we have in mind, the 2D case is sufficient.

We have the discrete convex optimization problem

$$\begin{aligned} \min \quad & f(p, \rho, u) = \left\langle A_{1x}(p_x^2) + A_{1y}(p_y^2), A_2(1/\rho) + a \right\rangle h_x h_y h_t \\ & + \left\langle u^2, A_2\left(1/(\mathbb{D}_2^T \rho) + 1/(\mathbb{D}_1^T \rho)\right) + c \right\rangle \gamma h_x h_y h_t \\ \text{subject to} \quad & D_{1x} p_x + D_{1y} p_y + D_2 \rho + D_3 u = b. \end{aligned}$$

The Lagrangian of this problem is

$$\mathcal{L}(p, \rho, u) = f(p, \rho, u) / (h_x h_y h_t) + \langle \lambda, D_{1x} p_x + D_{1y} p_y + D_2 \rho + D_3 u - b \rangle.$$

In the above,

$$a_{i,j,k} = \begin{cases} 1/\rho_{i,j}^0/2 & \text{if } k = 1, \\ 1/\rho_{i,j}^1/2 & \text{if } k = n_t, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\begin{aligned} b_{i,j,k} &= \begin{cases} \rho_{i,j}^0/h_t & \text{if } k = 1, \\ -\rho_{i,j}^1/h_t & \text{if } k = n_t, \\ 0 & \text{otherwise,} \end{cases} \\ c_{i,j,k} &= \begin{cases} 1/\mathbb{D}_2^T \rho_{i,j}^0/2 + 1/\mathbb{D}_1^T \rho_{i,j}^0/2 & \text{if } k = 1, \\ 1/\mathbb{D}_2^T \rho_{i,j}^1/2 + 1/\mathbb{D}_1^T \rho_{i,j}^1/2 & \text{if } k = n_t, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The KKT conditions now are

$$\begin{aligned} \nabla_{p_x} \mathcal{L} &= D_{1x}^T \lambda + 2p_x \circ A_{1x}^T (A_2(1/\rho) + a) = 0 \\ \nabla_{p_y} \mathcal{L} &= D_{1y}^T \lambda + 2p_y \circ A_{1y}^T (A_2(1/\rho) + a) = 0 \\ \nabla_{\rho} \mathcal{L} &= D_2^T \lambda - A_2^T (A_{1x}(p_x^2) + A_{1y}(p_y^2)) / \rho^2 - \gamma \mathbb{D}_2 \left(A_2^T(u^2) / (\mathbb{D}_2^T \rho)^2 \right) \\ &\quad - \gamma \mathbb{D}_1 \left(A_2^T(u^2) / (\mathbb{D}_1^T \rho)^2 \right) = 0 \\ \nabla_u \mathcal{L} &= D_3^T \lambda + 2\gamma u \circ \left(A_2 \left(1/(\mathbb{D}_2^T \rho) + 1/(\mathbb{D}_1^T \rho) \right) + c \right) = 0 \\ \nabla_{\lambda} \mathcal{L} &= D_{1x} p_x + D_{1y} p_y + D_2 \rho + D_3 u - b = 0, \end{aligned}$$

with $\circ$ denoting block-wise multiplication.

Let $w = (p_x, p_y, \rho, u)$, then at each SQP iteration we solve the system

$$\begin{pmatrix} \hat{A} & D^* \\ D & 0 \end{pmatrix} \begin{pmatrix} \delta w \\ \delta \lambda \end{pmatrix} = - \begin{pmatrix} \nabla_w \mathcal{L} \\ \nabla_{\lambda} \mathcal{L} \end{pmatrix}, \quad (19)$$

where $D = (D_{1x}, D_{1y}, D_2, D_3)$. The matrix $\hat{A}$ is an approximation of the Hessian of the objective function

$$\begin{pmatrix} 2\text{diag}(A_{1x}^T (A_2 (1/\rho) + a)) & 0 \\ 0 & 2\text{diag}(A_{1y}^T (A_2 (1/\rho) + a)) \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \text{diag}(g(p, \rho, u)) & 0 \\ 0 & 2\gamma \text{diag}(A_2 (1/(\mathbb{D}_2^T \rho) + 1/(\mathbb{D}_1^T \rho)) + c) \end{pmatrix}.$$

The operator $g(p, \rho, u)$ is the Hessian of f over ρ with $g_{i,j,k+\frac{1}{2}}$ being the map

$$\begin{aligned} g_{i,j+\frac{1}{2}}(X) = & 2 \left(A_2^T \left(A_{1x} (p_x^2) + A_{1y} (p_y^2) \right) \right)_{i,j+\frac{1}{2}} / \rho_{i,j+\frac{1}{2}}^3 X \\ & + 2\gamma \mathbb{D}_2 \left[\left(A_2^T (u^2) \right)_{i,j+\frac{1}{2}} / \left(\mathbb{D}_2^T \rho \right)_{i,j+\frac{1}{2}}^3 \mathbb{D}_2^T X \right] \\ & + 2\gamma \mathbb{D}_1 \left[\left(A_2^T (u^2) \right)_{i,j+\frac{1}{2}} / \left(\mathbb{D}_1^T \rho \right)_{i,j+\frac{1}{2}}^3 \mathbb{D}_1^T X \right]. \end{aligned}$$

References

1. Ambrosio, L., Gigli, N., Savaré, G.: Gradient Flows in Metric Spaces and in the Space of Probability Measures. Springer, Berlin (2006)
2. Angenent, S., Haker, S., Tannenbaum, A.: Minimizing flows for the Monge–Kantorovich problem. *SIAM J. Math. Anal.* **35**(1), 61–97 (2003)
3. Ascher, U.M.: Numerical Methods for Evolutionary Differential Equations. SIAM, Philadelphia (2008)
4. Benamou, J.-D., Brenier, Y.: A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numer. Math.* **84**(3), 375–393 (2000)
5. Benamou, J.-D., Froese, B.D., Oberman, A.M.: Numerical solution of the optimal transportation problem using the Monge–Ampère equation. *J. Comput. Phys.* **260**, 107–126 (2014)
6. Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., Peyré, G.: Iterative Bregman projections for regularized transportation problems. *SIAM J. Sci. Comput.* **37**(2), A1111–A1138 (2015)
7. Brenier, Y.: Polar factorization and monotone rearrangement of vector-valued functions. *Commun. Pure Appl. Math.* **44**(4), 375–417 (1991)
8. Byrd, R.H., Curtis, F.E., Nocedal, J.: An inexact SQP method for equality constrained optimization. *SIAM J. Optim.* **19**(1), 351–369 (2008)
9. Carlen, E.A., Maas, J.: Gradient flow and entropy inequalities for quantum markov semigroups with detailed balance. *arXiv preprint arXiv:1609.01254*, (2016)
10. Chen, Y., Georgiou, T.T., Pavon, M.: Entropic and displacement interpolation: a computational approach using the Hilbert metric. *arXiv:1506.04255v1* (2015)
11. Chen, Y., Georgiou, T. T., Tannenbaum, A.: Matrix optimal mass transport: a quantum mechanical approach. *arXiv preprint arXiv:1610.03041* (2016)
12. Chen, Y., Georgiou, T.T., Tannenbaum, A.: Vector-valued optimal mass transport. *arXiv preprint arXiv:1611.09946* (2016)
13. Chen, Y., Georgiou, T.T., Tannenbaum, A.: Interpolation of density matrices and matrix-valued measures: the unbalanced case. *arXiv preprint arXiv:1612.05914* (2016)
14. Chow, S.-N., Dieci, L., Li, W., Zhou, H.: Entropy dissipation semi-discretization schemes for Fokker–Planck equations. *arxiv:1608.02628* (2017)
15. Cuturi, M.: Sinkhorn distances: lightspeed computation of optimal transport. In: Proceedings of the Advances in Neural Information Processing Systems (NIPS), pp. 2292–2300 (2013)

16. Erbar, M., Maas, J.: Ricci curvature of finite Markov chains via convexity of the entropy. *Arch. Ration. Mech. Anal.* **206**, 997–1038 (2012)
17. Gangbo, W., McCann, R.J.: The geometry of optimal transportation. *Acta Math.* **177**(2), 113–161 (1996)
18. Haber, E., Horesh, R.: A multilevel method for the solution of time dependent optimal transport. *Numer. Math. Theory Methods Appl.* **8**(01), 97–111 (2015)
19. Jordan, R., Kinderlehrer, D., Otto, F.: The variational formulation of the Fokker–Planck equation. *SIAM J. Math. Anal.* **29**(1), 1–17 (1998)
20. Kantorovich, L.V.: On the transfer of masses. *Dokl. Akad. Nauk. SSSR* **37**(7–8), 227–229 (1942)
21. Kershaw, D.S.: The incomplete Cholesky conjugate gradient method for the iterative solution of systems of linear equations. *J. Comput. Phys.* **26**(1), 43–65 (1978)
22. Li, W., Yin, P., Osher, S.: A fast algorithm for unbalanced L_1 Monge–Kantorovich problem. *CAM Report*, pp. 16–84 (2016)
23. Li, W., Yin, P., Osher, S.: Computations of optimal transport distance with Fisher information regularization. *J. Sci. Comput.* (2018) (To appear in) [arxiv:1704.04605](https://arxiv.org/abs/1704.04605)
24. Li, W., Ryu, E.K., Osher, S., Yin, W., Gangbo, W.: A parallel method for earth mover’s distance. *J. Sci. Comput.* **75**, 1–16 (2017)
25. McCann, R.J.: A convexity principle for interacting gases. *Adv. Math.* **128**(1), 153–179 (1997)
26. Mittnenzweig, M., Mielke, A.: An entropic gradient structure for Lindblad equations and GENERIC for quantum systems coupled to macroscopic models. *arXiv preprint* [arXiv:1609.05765](https://arxiv.org/abs/1609.05765) (2016)
27. Monge, G.: *Mémoire sur la Théorie des Déblais et des Remblais*. De l’Imprimerie Royale, Paris, France (1781)
28. Ning, L., Georgiou, T.T., Tannenbaum, A.: On matrix-valued Monge–Kantorovich optimal mass transport. *IEEE Trans. Autom. Control* **60**(2), 373–382 (2015)
29. Nocedal, J., Wright, S.: *Numerical Optimization*. Springer, Berlin (2006)
30. Otto, F., Villani, C.: Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.* **173**(2), 361–400 (2000)
31. Rachev, S.T., Rüschendorf, L.: *Mass Transportation Problems: Volume I—Theory*. Springer, Berlin (1998)
32. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.* **14**(5), 877–898 (1976)
33. Ryu, E.K., Chen, Y., Li, W., Osher, S.: Vector and matrix optimal mass transport: theory, algorithm, and applications. [arxiv:1712.10279](https://arxiv.org/abs/1712.10279) (2017)
34. Steklova, K., Haber, E.: Joint hydrogeophysical inversion: state estimation for seawater intrusion models in 3D. *Comput. Geosci.* **21**(1), 75–94 (2017)
35. Tannenbaum, A., Sander, C., Zhu, L., Sandhu, R., Kolesov, I., Reznik, E., Senbabaoglu, Y., Georgiou, T.: Ricci curvature and robustness of cancer networks. [arxiv:1502.04512](https://arxiv.org/abs/1502.04512) (2015)
36. Villani, C.: *Topics in Optimal Transportation*, vol. 58. American Mathematical Society, Providence (2003)