A PARALLEL METHOD FOR EARTH MOVER'S DISTANCE

WUCHEN LI, ERNEST K. RYU, STANLEY OSHER, WOTAO YIN, AND WILFRID GANGBO

Abstract. We propose a new algorithm to approximate the Earth Mover's distance (EMD). Our main idea is motivated by the theory of optimal transport, in which EMD can be reformulated as a familiar L_1 type minimization. We use a regularization which gives us a unique solution for this L_1 type problem. The new regularized minimization is very similar to problems which have been solved in the fields of compressed sensing and image processing, where several fast methods are available. In this paper, we adopt a primal-dual algorithm designed there, which uses very simple updates at each iteration and is shown to converge very rapidly. Several numerical examples are provided.

1. Introduction

The earth mover's distance (EMD) has been used extensively in fields such as image processing, computer vision and statistics [13, 15, 27, 17]. E.g. EMD has been widely used in image retrieval problems [21]. In this paper, we present a new method to approximate the EMD. This method is simple to implement and simple to parallelize.

We begin by reviewing the definitions and basic results relating to EMD. Let $\Omega \subset \mathbb{R}^d$ be convex and compact and let $c : \Omega \times \Omega \to [0, \infty)$ be a distance function in Ω. For any pair of non-negative measures ρ^0, ρ^1 on Ω with equal mass, EMD is defined by the minimization problem

$$\mathbf{EMD}(\rho^0, \rho^1) = \left(\begin{array}{ll} \text{minimize} & \int_{\Omega\times\Omega} c(\mathbf{x}_1, \mathbf{x}_2)\pi(\mathbf{x}_1, \mathbf{x}_2)\, d\mathbf{x}_1 d\mathbf{x}_2 \\ \text{subject to} & \int_\Omega \pi(\mathbf{x}_1, \mathbf{x}_2)\, d\mathbf{x}_2 = \rho^0(\mathbf{x}_1) \\ & \int_\Omega \pi(\mathbf{x}_1, \mathbf{x}_2)\, d\mathbf{x}_1 = \rho^1(\mathbf{x}_2) \end{array} \right), \tag{1}$$

where $\pi \geq 0$, a joint measure (transport plan) on $\Omega \times \Omega$, is the optimization variable. Note that $\pi(\mathbf{x}_1, \mathbf{x}_2)$ is constrained to have $\rho^0(\mathbf{x}_1)$ and $\rho^1(\mathbf{x}_2)$ as its marginals.

We call the distance function c the *ground metric*. The domain Ω and the ground metric c define the EMD. In this paper, we use the Euclidean distance (L_2) [3, 4] and the Manhattan distance (L_1) [14] for the ground metric. They correspond to, respectively, $c(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_2$ and $c(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_1$. We call (1) with the L_1, L_2 ground metric the EMD-L_1 and EMD-L_2 problems, respectively.

In recent years, the optimization problem (1) has been studied extensively in the field of optimal transport [8, 24, 28]. Many interesting metrics, including the Euclidean and

Key words and phrases. Earth Mover's distance; Optimal transport; Compressed sensing; Primal-dual algorithm; L_1 regularization.

This work is partially supported by ONR grants N000141410683, N000141210838 and DOE grant DE-SC00183838.

Manhattan distances, can be represented in the variational form

$$c(\mathbf{x}_1, \mathbf{x}_2) = \begin{pmatrix} \text{minimize} & \int_0^1 L(\mathbf{v}(t))\, dt \\ \text{subject to} & \frac{d}{dt}\mathbf{x} = \mathbf{v}\ , \quad \mathbf{x}(0) = \mathbf{x}_1\ , \quad \mathbf{x}(1) = \mathbf{x}_2 \end{pmatrix}\ ,$$

where the infimum is taken among all continuous differentiable path $\gamma(t) \in \Omega$ and the Lagrangian, $L(\mathbf{v})$, is homogeneous of degree 1 and convex in $\mathbf{v}$. For example, $L(\mathbf{v}) = \|\mathbf{v}\|_2$ yields the Euclidean distance, and $L(\mathbf{v}) = \|\mathbf{v}\|_1$ the Manhattan distance. When this is the case, remarkably, EMD can be equivalently written as

$$\mathbf{EMD}(\rho^0, \rho^1) = \begin{pmatrix} \text{minimize} & \int_\Omega L(\mathbf{m}(\mathbf{x}))\, d\mathbf{x} \\ \text{subject to} & \nabla \cdot \mathbf{m}(\mathbf{x}) + \rho^1(\mathbf{x}) - \rho^0(\mathbf{x}) = 0 \\ & \mathbf{m}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) = 0, \text{ for all } \begin{cases} \mathbf{x} \in \partial\Omega, \\ \mathbf{n}(\mathbf{x}) \text{ normal to } \partial\Omega \end{cases} \end{pmatrix} \tag{2}$$

where the optimization variable $\mathbf{m} : \Omega \to \mathbb{R}^d$ is a flux vector satisfying the zero flux boundary condition [1, 4]. The connection between (1) and (2) is briefly explained in section 2.

The formulation (2) has huge computational benefits. First, the size of the optimization variable in (2) is much smaller than that of (1), when solving a discrete approximation; when using a discretized grid with N points, the variable size is reduced from N^2 to N. Second, (2) is an L_1-type minimization problem, which shares its structure with many problems in compressed sensing and image processing, and therefore we can take advantage of well-established fast and simple algorithms [10, 22, 29].

In this paper, we propose a new algorithm to compute the EMD that leverages the structure of the formulation (2), which, roughly speaking, has the form

$$\begin{aligned} \mathbf{m}^{k+1} &= \text{shrink}(\mathbf{m}^k + \mu \nabla \Phi^k) \\ \Phi^{k+1} &= \Phi^k + \tau(\text{div}(2\mathbf{m}^{k+1} - \mathbf{m}^k) + \rho^1 - \rho^0)\ . \end{aligned}$$

Here $\mu, \tau > 0$ are the algorithm's parameters, ∇, div are discrete gradient, divergence operators respectively, and the *shrink operator* shrink$(\cdot)$ is a simple function that depends on the ground metric. Under appropriate conditions, $\mathbf{m}^k$ converge to a solution, and Φ^k converge to a Lagrange multiplier. The algorithm discretizes the domain Ω with a finite volume approximation and then applies the first-order primal-dual method of Chambolle and Pock for the optimization [6, 18]. This method is very simple to implement and, as we discuss, very easy to parallelize.

To compute the EMD, algorithms based on linear programming [11, 14] and the alternating direction method of multipliers (ADMM) [3, 4, 27] have been proposed. Compared to these existing methods, our method has a much lower computational cost per iteration (though it can take more iterations to converge) because *no* linear system (in particular, no elliptic problem) is solved at each iteration. Our method is very simple, and this simplicity makes the method easy to parallelize. We implemented our algorithm with CUDA C++ and run it on a GPU. Its performance is presented in Section 5.

Besides, proximal splitting methods have also been applied to some optimal transport related minimization problems [5, 16], in which the Lagrangian L is not homogeneous of degree 1 and the formulation (2) is time dependent. One of the hardest problem there

is to handle the non-negativity of density functions in each time level. However, there is not such an issue in EMD-L_1 or EMD-L_2 computation. This is because the optimization problem is static and the shrink operator is simple. The proposed algorithm is also robust, in the sense of handling various measures. It is especially true for ρ^0, ρ^1 being sparse, such as delta measures.

The rest of this paper is organized as follows. We provide a short review on EMD in section 2 and describe the proposed algorithm in section 3. Several parallel computational considerations and numerical examples are discussed in sections 4 and 5, respectively. We make conclusions in section 6.

2. Review of Optimal transport

For the reader's convenience, we provide a short review on the equivalence between (1) and (2). The connection can be derived in two ways. In the first way, (2) is derived as the bi-dual (dual of the dual) to the linear program (1); see [1, 4, 28] for details. The other way is based on an optimal control viewpoint, which we discuss. Along with this, we briefly summarize the history of optimal transport.

In 1781, Monge first proposed the problem of optimal transport:

$$\begin{array}{ll} \text{minimize} & \int_\Omega c(\mathbf{x}_1, T(\mathbf{x}_1))\rho^0(\mathbf{x}_1)\, d\mathbf{x}_1 \\ \text{subject to} & \rho^1(T(\mathbf{x}_1))\det(\nabla T(\mathbf{x}_1)) = \rho^0(\mathbf{x}_1)\ , \end{array} \tag{3}$$

where the minimization variable is the map T, a one-to-one smooth mapping that transfers ρ^0 to ρ^1. Because T is possibly nonlinear, the optimization problem (3) is generally nonlinear. In the 1940s Kantorovich identified that (3) can be solved with the linear program (1). Today, it is known that under suitable conditions on ρ^0 and ρ^1, the minimal values of (3) and (1) are identical and the minimizing joint measure π of (1) exists. From its support, one can find the optimal map T.

(3) has an important reformation, which connects to optimal control [1, 2, 3]. By writing c in a variational form, i.e.,

$$c(\mathbf{x}_1, T(\mathbf{x}_1)) = \left(\begin{array}{ll} \text{minimize} & \int_0^1 L(\mathbf{v})\, dt \\ \text{subject to} & \frac{d}{dt}\mathbf{x} = \mathbf{v}\ , \quad \mathbf{x}(0) = \mathbf{x}_1\ , \quad \mathbf{x}(1) = T(\mathbf{x}_1)\ , \end{array} \right)$$

we can reformulate (3) as

$$\begin{array}{ll} \text{minimize} & \int_\Omega \int_0^1 L(\mathbf{v})\rho(t,\mathbf{x})\, dt d\mathbf{x} \\ \text{subject to} & \frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\mathbf{v}) = 0 \\ & \rho(0,\mathbf{x}) = \rho^0\ , \quad \rho(1,\mathbf{x}) = \rho^1\ , \end{array} \tag{4}$$

where the minimum is taken among all Borel vector fields $\mathbf{v}(t,\mathbf{x})$ (satisfying the zero flux condition on $\partial\Omega$) and density function $\rho(t,x)$ that transports ρ^0 to ρ^1 continuously in time. The minimization problem (4) is just the dynamical version of (3), and the optimal map can be obtained through

$$T(\mathbf{x}_1) = \mathbf{x}(1)\ ,$$

where $\mathbf{x}(t)$ solves the following initial value ordinary differential equation (ODE) [8]:

$$\frac{d}{dt}\mathbf{x} = \mathbf{v}(t,\mathbf{x}(t))\ , \quad \mathbf{x}(0) = \mathbf{x}_1\ . \tag{5}$$

If L is homogeneous degree 1 and convex in $\mathbf{v}$ (think, for example, $L(\mathbf{x},\mathbf{v}) = \|\mathbf{v}\|_2$) [28], then (4) is equivalent to the time independent (static) minimization problem (2). Given an $\mathbf{m}$ feasible for (2), define $\rho(t,\mathbf{x}) = t\rho^1(\mathbf{x}) + (1-t)\rho^0(\mathbf{x})$ and $\mathbf{v}(t,\mathbf{x}) = \mathbf{m}(\mathbf{x})/\rho(t,\mathbf{x})$. Then $\mathbf{v}(t,\mathbf{x})$ is feasible for (4) and has the same objective value as $\mathbf{m}$ did for (2). So

$$\inf_{\mathbf{v}} \int_\Omega \int_0^1 L(\mathbf{v})\rho(t,\mathbf{x})\, dtd\mathbf{x} \leq \inf_{\mathbf{m}} \int_\Omega L(\mathbf{m}(\mathbf{x}))\, d\mathbf{x}\ ,$$

The other direction follows from Jensen's inequality:

$$\int_0^1 L(\mathbf{v})\rho(t,\mathbf{x})\, dt \geq L\left(\int_0^1 \mathbf{v}\rho(t,\mathbf{x})\, dt\right) = L(\mathbf{m}(\mathbf{x}))\ ,$$

where

$$\mathbf{m}(\mathbf{x}) = \int_0^1 \rho(t,\mathbf{x})\mathbf{v}(t,\mathbf{x})\, dt\ .$$

So

$$\inf_{\mathbf{v}} \int_\Omega \int_0^1 L(\mathbf{v})\rho(t,\mathbf{x})\, d\mathbf{x}dt \geq \inf_{\mathbf{m}} \int_\Omega L(\mathbf{m}(\mathbf{x}))\, d\mathbf{x}\ ,$$

and we conclude (4) and (2) have the same optimal value.

In conclusion, the four minimization problems (1), (2), (3), and (4) are equivalent, and they share the same minimal value. In this paper, we focus on (2) for efficient computation.

3. Proposed Algorithm

The EMD problem, as presented in (2), has similar structures to many homogeneous degree one regularized problems. In this section we use a finite volume discretization to approximate (2). The discretized problem becomes an L_1-type optimization with linear constraints, which allows us to apply the hybrid primal-dual method designed in [6, 18].

3.1. Discretization.

For notational simplicity, we will consider the case where $\Omega \subset \mathbb{R}^2$ and Ω is square. The following discussion does immediately generalize to higher dimensions and more complicated domains.

Also, we will use the same symbol to denote the discretizations and their continuous counterparts. Whether we are referring to the continuous variable or its discretization should be clear from the context.

Consider a $n \times n$ discretization of Ω with finite difference Δx in both x and y directions. Write the x and y coordinates of the points as $x_1,\ldots,x_n$ and $y_1,\ldots,y_n$. So we are approximating the domain Ω with $\{x_1,\ldots,x_n\} \times \{y_1,\ldots,y_n\}$. Write $C(x,y)$ be the $\Delta x \times \Delta x$ cube centered at (x,y), i.e.,

$$C(x,y) = \{(x',y') \in \mathbb{R}^2 \,|\, |x'-x| \leq \Delta x/2\,,\ |y'-y| \leq \Delta x/2\}\ .$$

We use a finite volume approximation for ρ^0 and ρ^1. Specifically, we write $\rho^0 \in \mathbb{R}^{n\times n}$ with

$$\rho^0_{ij} \approx \int_{C(x_i,y_j)} \rho^0(x,y)\, dxdy\ ,$$

for $i,j = 1,\ldots,n$. The discretization $\rho^1 \in \mathbb{R}^{n\times n}$ is defined the same way.

Write $\mathbf{m} = (m_x, m_y)$ for both the continuous variable and its discretization. To be clear, the subscripts of m_x and m_y do not denote differentiation. We use the discretization $m_x \in \mathbb{R}^{(n-1)\times n}$ and $m_y \in \mathbb{R}^{n\times(n-1)}$. For $i = 1, \ldots, n-1$ and $j = 1, \ldots, n$

$$m_{x,ij} \approx \int_{C(x_i+\Delta x/2, y_j)} m_x(x,y)\, dxdy \ ,$$

and for $i = 1, \ldots, n$ and $j = 1, \ldots, n-1$

$$m_{y,ij} \approx \int_{C(x_i, y_j+\Delta x/2)} m_y(x,y)\, dxdy \ .$$

In defining m_x and m_y, the center points are placed between the $n \times n$ grid points to make the finite difference operator symmetric.

Define the discrete divergence operator $\mathrm{div}(\mathbf{m}) \in \mathbb{R}^{n\times n}$ as

$$\mathrm{div}(\mathbf{m})_{ij} = \frac{1}{\Delta x}(m_{x,ij} - m_{x,(i-1)j} + m_{y,ij} - m_{y,i(j-1)}) \ ,$$

for $i, j = 1, \ldots, n$, where we mean $m_{x,0j} = m_{x,nj} = 0$ for $j = 1, \ldots, n$ and $m_{y,i0} = m_{y,in} = 0$ for $i = 1, \ldots, n$. This definition of $\mathrm{div}(\mathbf{m})$ makes the discrete approximation be consistent with the zero-flux boundary condition.

For $\Phi \in \mathbb{R}^{n\times n}$, define the discrete gradient operator $\nabla\Phi = ((\nabla\Phi)_x, (\nabla\Phi)_y)$ as

$$\begin{aligned}(\nabla\Phi)_{x,ij} &= (1/\Delta x)\,(\Phi_{i+1,j} - \Phi_{i,j}) \quad \text{for } i = 1, \ldots, n-1,\ j = 1, \ldots, n \\ (\nabla\Phi)_{y,ij} &= (1/\Delta x)\,(\Phi_{i,j+1} - \Phi_{i,j}) \quad \text{for } i = 1, \ldots, n,\ j = 1, \ldots, n-1 \ .\end{aligned}$$

So $(\nabla\Phi)_x \in \mathbb{R}^{(n-1)\times n}$ and $(\nabla\Phi)_y \in \mathbb{R}^{n\times(n-1)}$, and the ∇ is the adjoint of $-\mathrm{div}$.

We will soon see that using ghost cells is convenient for both describing and implementing the method. So we define the variable $\tilde{\mathbf{m}} = (\tilde{m}_x, \tilde{m}_y) \in \mathbb{R}^{2\times n\times n}$ where

$$\tilde{m}_{x,ij} = \begin{cases} m_{x,ij} & \text{for } i < n \\ 0 & \text{for } i = n \end{cases}$$
$$\tilde{m}_{y,ij} = \begin{cases} m_{y,ij} & \text{for } j < n \\ 0 & \text{for } j = n \ , \end{cases}$$

for $i, j = 1, \ldots, n$. We also define $\tilde{\nabla}\Phi = ((\tilde{\nabla}\Phi)_x, (\tilde{\nabla}\Phi)_y) \in \mathbb{R}^{2\times n\times n}$, where

$$(\tilde{\nabla}\Phi)_{x,ij} = \begin{cases} (\nabla\Phi)_{x,ij} & \text{for } i < n \\ 0 & \text{for } i = n \end{cases}$$
$$(\tilde{\nabla}\Phi)_{y,ij} = \begin{cases} (\nabla\Phi)_{y,ij} & \text{for } j < n \\ 0 & \text{for } j = n \ , \end{cases}$$

for $i, j = 1, \ldots, n$. Finally, we write $\tilde{\mathbf{m}} = (\tilde{m}_x, \tilde{m}_y)$ and $\tilde{\mathbf{m}}_{ij} = (\tilde{m}_{x,ij}, \tilde{m}_{y,ij})$ and $(\nabla\Phi)_{ij} = ((\nabla\Phi)_{x,ij}, (\nabla\Phi)_{y,ij})$ for $i, j = 1, \ldots, n$

3.2. **EMD with L_2 ground metric.** Using this notation, we write the discretization of (2) as

$$\begin{aligned} \underset{\mathbf{m}}{\text{minimize}} \quad & \|\mathbf{m}\|_{1,2} \\ \text{subject to} \quad & \mathrm{div}(\mathbf{m}) + \rho^1 - \rho^0 = 0 \ , \end{aligned} \tag{6}$$

where $m_x \in \mathbb{R}^{(n-1)\times n}$ and $m_y \in \mathbb{R}^{n\times(n-1)}$ are the optimization variables. The boundary conditions implicitly handled by the discretization. The objective is

$$\|\mathbf{m}\|_{1,2} = \sum_{i=1}^{n}\sum_{j=1}^{n} \|\mathbf{m}_{ij}\|_2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \sqrt{m_{x,ij}^2 + m_{y,ij}^2} \ ,$$

where we mean $m_{x,nj} = 0$ for $j = 1, \ldots, n$ and $m_{y,in} = 0$ for $i = 1, \ldots, n$.

Define the Lagrangian

$$L(\mathbf{m}, \Phi) = \|\mathbf{m}\|_{1,2} + \langle \Phi, \operatorname{div}(\mathbf{m}) + \rho^1 - \rho^0 \rangle \ ,$$

where $\Phi \in \mathbb{R}^{n\times n}$ is the Lagrange multiplier corresponding to the equality constraint of (6). Here $\langle \cdot, \cdot \rangle$ denotes the inner product between $n \times n$ matrices treated as vectors, i.e.,

$$\langle A, B \rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}B_{ij} \ .$$

Standard convex analysis states that $\mathbf{m}^\star$ is a solution to (6) if and only if there is a $\Phi^\star$ such that $(\mathbf{m}^\star, \Phi^\star)$ is a saddle point of $L(\mathbf{m}, \Phi)$ [19]. In other words, we can solve (6) by solving the minimax problem

$$\min_{\mathbf{m}} \max_{\Phi} \quad L(\mathbf{m}, \Phi). \tag{7}$$

Saddle point problems, such as (7), can be solved with the first-order primal-dual method of Chambolle and Pock [6, 18]:

$$\begin{aligned} \mathbf{m}^{k+1} &= \operatorname*{argmin}_{\mathbf{m}} \left\{ \|\mathbf{m}\|_{1,2} + \langle \Phi^k, \operatorname{div}(\mathbf{m}) \rangle + \frac{1}{2\mu}\|\mathbf{m} - \mathbf{m}^k\|_2^2 \right\} \\ \Phi^{k+1} &= \operatorname*{argmax}_{\Phi} \left\{ \left\langle \Phi, \operatorname{div}(2\mathbf{m}^{k+1} - \mathbf{m}^k) + \rho^1 - \rho^0 \right\rangle - \frac{1}{2\tau}\|\Phi - \Phi^k\|_2^2 \right\} \end{aligned} \tag{8}$$

where $\mu, \tau > 0$ are step sizes. The meaning of $\|\cdot\|_2^2$ is standard:

$$\|\mathbf{m} - \mathbf{m}^k\|_2^2 = \sum_{i=1}^{n-1}\sum_{j=1}^{n}(m_{x,ij} - m_{x,ij}^k)^2 + \sum_{i=1}^{n}\sum_{j=1}^{n-1}(m_{y,ij} - m_{y,ij}^k)^2$$

and

$$\|\Phi - \Phi^k\|_2^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(\Phi_{ij} - \Phi_{ij}^k)^2 \ .$$

These steps can be interpreted as a gradient descent in the primal variable $\mathbf{m}$ and a gradient ascent in the dual variable Φ.

It turns out the optimization problems that define (8) have explicit formulas that are separable over the indices i, j.

$$\begin{aligned}
&\operatorname*{argmin}_{\mathbf{m}} \left\{ \|\mathbf{m}\|_{1,2} + \langle \Phi^k, \nabla \cdot \mathbf{m} \rangle + \frac{1}{2\mu} \|\mathbf{m} - \mathbf{m}^k\|_2^2 \right\} \\
&= \operatorname*{argmin}_{\mathbf{m}} \left\{ \sum_{ij} \left(\|\mathbf{m}_{ij}\|_{1,2} + \frac{1}{\Delta x} \Phi_{ij}^k (m_{x,ij} - m_{x,(i-1)j} + m_{y,ij} - m_{y,i(j-1)}) + \frac{1}{2\mu} \|\mathbf{m}_{ij} - \mathbf{m}_{ij}^k\|_2^2 \right) \right\} \\
&= \operatorname*{argmin}_{\mathbf{m}} \left\{ \sum_{ij} \left(\|\mathbf{m}_{ij}\|_{1,2} - (\nabla \Phi^k)_{ij}^T \mathbf{m}_{ij} + \frac{1}{2\mu} \|\mathbf{m}_{ij} - \mathbf{m}_{ij}^k\|_2^2 \right) \right\},
\end{aligned}$$

where again, all out of bounds indicies are interpreted as zeros. This minimization has a closed form solution, which can be written concisely with $\tilde{\mathbf{m}}$ and $\tilde{\nabla}$:

$$\tilde{\mathbf{m}}_{ij}^{k+1} = \text{shrink}_2(\tilde{\mathbf{m}}_{ij}^k + \mu(\tilde{\nabla}\Phi^k)_{ij}, \mu)$$

for $i, j = 1, \ldots, n$. The shrink operator shrink_2 is defined as

$$\text{shrink}_2(v, \mu) = \begin{cases} (1 - \mu/\|v\|_2)v & \text{for } \|v\|_2 \geq \mu \\ 0 & \text{for } \|v\|_2 < \mu \end{cases}.$$

Note that shrink_2 maps from $\mathbb{R}^2$ to $\mathbb{R}^2$, given a fixed μ.

Likewise, we have

$$\begin{aligned}
&\operatorname*{argmax}_{\Phi} \left\{ \left\langle \Phi, \text{div}(2\mathbf{m}^{k+1} - \mathbf{m}^k) + \rho^1 - \rho^0 \right\rangle - \frac{1}{2\tau} \|\Phi - \Phi^k\|_2^2 \right\} \\
&= \operatorname*{argmax}_{\Phi} \left\{ \sum_{ij} \left(\Phi_{ij}((\text{div}(2\mathbf{m}^{k+1} - \mathbf{m}^k))_{ij} + \rho_{ij}^1 - \rho_{ij}^0) - \frac{1}{2\tau}(\Phi_{ij} - \Phi_{ij}^k)^2 \right) \right\},
\end{aligned}$$

and second line of (8) simplifies to

$$\Phi_{ij}^{k+1} = \Phi_{ij}^k + \tau((\text{div}(2\mathbf{m}^{k+1} - \mathbf{m}^k))_{ij} + \rho_{ij}^1 - \rho_{ij}^0)$$

for $i, j = 1, \ldots, n$.

We are now ready to state our algorithm.

Primal-Dual for EMD-L_2

Input: Discrete probabilities ρ^0, ρ^1
Initial guess of $\mathbf{m}^0$, step size μ, τ
Output: $\mathbf{m}$ and EMD value $\|\mathbf{m}\|_{1,2}$

---

1. **for** $k = 1, 2, \cdots$ (Iterate until convergence)
2. $\quad \tilde{\mathbf{m}}_{ij}^{k+1} = \text{shrink}_2(\tilde{\mathbf{m}}_{ij}^k + \mu(\tilde{\nabla}\Phi^k)_{ij}, \mu)$ for $i, j = 1, \ldots, n$
3. $\quad \Phi_{ij}^{k+1} = \Phi_{ij}^k + \tau((\text{div}(2\mathbf{m}^{k+1} - \mathbf{m}^k))_{ij} + \rho_{ij}^1 - \rho_{ij}^0)$ for $i, j = 1, \ldots, n$
4. **end**

Again, $\tilde{\mathbf{m}}$ and $\tilde{\nabla}\Phi$ correspond to $\mathbf{m}$ and $\nabla\Phi$ padded with ghost cells, as discussed in Section 3.1.

3.3. **EMD with L_1 ground metric.** We next consider EMD-L_1. The arguments and notation are similar as before, so we only outline the difference.

We write the discretization of (2) as

$$\begin{aligned} &\underset{\mathbf{m}}{\text{minimize}} && \|\mathbf{m}\|_{1,1} \\ &\text{subject to} && \operatorname{div}(\mathbf{m}) + \rho^1 - \rho^0 = 0 \ . \end{aligned} \tag{9}$$

The objective is

$$\|\mathbf{m}\|_{1,1} = \sum_{i=1}^{n}\sum_{j=1}^{n} \|\mathbf{m}_{ij}\|_1 = \sum_{i=1}^{n}\sum_{j=1}^{n} |m_{x,ij}| + |m_{y,ij}| \ ,$$

where we mean $m_{x,nj} = 0$ for $j = 1, \ldots, n$ and $m_{y,in} = 0$ for $i = 1, \ldots, n$.

(9) is an L_1 optimization problem with a convex objective function and linear constraints. However, (9) can have multiple minimizers as the objective function is not strictly convex. To remedy this issue, we add quadratic regularization with a small $\epsilon > 0$,

$$\begin{aligned} &\underset{\mathbf{m}}{\text{minimize}} && \|\mathbf{m}\|_{1,1} + (\epsilon/2)\|\mathbf{m}\|_2^2 \\ &\text{subject to} && \operatorname{div}(\mathbf{m}) + \rho^1 - \rho^0 = 0 \ . \end{aligned} \tag{10}$$

Since its objective function is strictly convex, (10) does have a unique solution. It is worth mentioning our algorithm can still solve (2) without the regularization term and obtain one of its possibly many solutions.

As before, define the Lagrangian

$$L(\mathbf{m}, \Phi) = \|\mathbf{m}\|_{1,1} + (\epsilon/2)\|\mathbf{m}\|_2^2 + \langle \Phi, \operatorname{div}(\mathbf{m}) + \rho^1 - \rho^0 \rangle \ .$$

Again, we can solve (10) by solving

$$\min_{\mathbf{m}} \max_{\Phi} L(\mathbf{m}, \Phi) \ . \tag{11}$$

Again, we find a saddle point of (11) by the first order primal-dual algorithm [6, 18]

$$\begin{aligned} \mathbf{m}^{k+1} &= \underset{\mathbf{m}}{\operatorname{argmin}} \left\{ \|\mathbf{m}\|_{1,1} + (\epsilon/2)\|\mathbf{m}\|_2^2 + \langle \Phi^k, \nabla \cdot \mathbf{m} \rangle + \frac{1}{2\mu}\|\mathbf{m} - \mathbf{m}^k\|_2^2 \right\} \\ \Phi^{k+1} &= \underset{\Phi}{\operatorname{argmax}} \left\{ \left\langle \Phi, \operatorname{div}(2\mathbf{m}^{k+1} - \mathbf{m}^k) + \rho^1 - \rho^0 \right\rangle - \frac{1}{2\tau}\|\Phi - \Phi^k\|_2^2 \right\} \ . \end{aligned} \tag{12}$$

As in the EMD-L_2 setting, we have explicit formulas that are separable over the indices i, j for (12). The Φ update is the same as before, and the $\mathbf{m}$ update is

$$\begin{aligned} \tilde{m}_{x,ij}^{k+1} &= 1/(1+\epsilon\mu)\operatorname{shrink}_1(\tilde{m}_{x,ij}^k + \mu(\tilde{\nabla}\Phi^k)_{x,ij}, \mu) \\ \tilde{m}_{y,ij}^{k+1} &= 1/(1+\epsilon\mu)\operatorname{shrink}_1(\tilde{m}_{y,ij}^k + \mu(\tilde{\nabla}\Phi^k)_{y,ij}, \mu) \end{aligned}$$

for $i, j = 1, \ldots, n$, where shrink_1 operation is the shrink operator

$$\text{shrink}_1(v, \mu) = \begin{cases} (1 - \mu/|v|)v & \text{for } |v| \geq \mu \\ 0 & \text{for } |v| < \mu \ . \end{cases}$$

Note that shrink_1 maps from $\mathbb{R}$ to $\mathbb{R}$, given a fixed μ. The update for Φ^{k+1} is the same as before. Now we can write

---

Primal-dual method for EMD $- L_1$

Input: Discrete probabilities ρ^0, ρ^1;
Initial guess of $\mathbf{m}^0$, parameter $\epsilon > 0$, step size μ, τ.
Output: $\mathbf{m}$ and EMD value $\|\mathbf{m}\|_{1,1}$.

---

1. for $k = 1, 2, \cdots$ (Iterate until convergence)
2. $\quad \tilde{m}^{k+1}_{c,ij} = 1/(1 + \epsilon\mu)\text{shrink}_1(\tilde{m}^k_{c,ij} + \mu(\tilde{\nabla}\Phi^k)_{c,ij}, \mu)$ for $i, j = 1, \ldots, n$ and $c = x, y$
3. $\quad \Phi^{k+1}_{ij} = \Phi^k_{ij} + \tau((\text{div}(2\mathbf{m}^{k+1} - \mathbf{m}^k))_{ij} + \rho^1_{ij} - \rho^0_{ij})$ for $i, j = 1, \ldots, n$
4. **end**

Again, $\tilde{\mathbf{m}}$ and $\tilde{\nabla}\Phi$ correspond to $\mathbf{m}$ and $\nabla\Phi$ padded with ghost cells, as discussed in Section 3.1.

3.4. **Convergence analysis.** We now show that the proposed primal-dual algorithm converges to the minimizer of (6) and (10).

Define the discrete Laplacian operator as $\nabla^2 = \text{div} \cdot \nabla$.

Theorem 1. *Assume $\tau\mu < 1/\lambda_{\max}(\nabla^2)$, where $\lambda_{\max}(\nabla^2)$ denotes the largest eigenvalue of the discrete Laplacian operator ∇^2. Then with iterations* (8) *and* (12)

$$(\mathbf{m}^k, \Phi^k) \to (\mathbf{m}^\star, \Phi^\star) \ ,$$

where $(\mathbf{m}^\star, \Phi^\star)$ is a saddle point of L in (7) *or* (11). *Define*

$$R^k = (1/\mu)\|\mathbf{m}^{k+1} - \mathbf{m}^k\|_2^2 + (1/\tau)\|\Phi^{k+1} - \Phi^k\|_2^2 - 2\langle \Phi^{k+1} - \Phi^k, \text{div}(\mathbf{m}^{k+1} - \mathbf{m}^k)\rangle \ .$$

Then $R^k \geq 0$ and $R^k = 0$ if and only if $(\mathbf{m}^k, \Phi^k)$ is a saddle point of (7) *or* (11). *R^k monotonically converges to* 0.

Proof. We check the conditions required in [6]. Let us rewrite L by

$$L(\mathbf{m}, \Phi) = G(\mathbf{m}) + \Phi^T K\mathbf{m} - F(\Phi) \ ,$$

where $G(\mathbf{m}) = \|\mathbf{m}\|_{1,2}$ or $G(\mathbf{m}) = \|\mathbf{m}\|_{1,1} + (\epsilon/2)\|\mathbf{m}\|_2^2$, $K = \text{div}$, and $F(\Phi) = \sum_{ij} \Phi_{ij}(\rho^0_{ij} - \rho^1_{ij})$. Observe that G, F are convex functions and K is a linear operator. Since $\nabla^2 = KK^T$, the algorithm converges for $\mu\tau\|\nabla^2\|_2^2 < 1$.

The Chambolle-Pock methods can be interpreted as a proximal point method under a certain metric [12]. R^k is the fixed-point residual of the non-expansive mapping defined by the proximal point method and thus decreases monotonically to 0, c.f., review paper [23]. □

4. Computational considerations

Parallelizing the methods for EMD-L_2 and EMD-L_1 is simple. We can split the computation over the indices (i, j) as follows:

```
m_temp[i,j] = m[i,j]
m[i,j] = shrink(m[i,j]+mu/dx*(Phi[i+1,j]-Phi[i+1,j],Phi[i,j+1]-Phi[i+1,j]))
m_temp[i,j] = 2*m[i,j]-m_temp[i,j]
-------------------------------------------------------------------------
Synchronize over all i,j
-------------------------------------------------------------------------
divm[i,j] = m_temp_x[i,j]-m_temp_x[i-1,j]+m_temp_y[i,j]-m_temp_y[i,j-1]
Phi[i,j] = Phi[i,j] + tau*(divm[i,j]/dx+rho1[i,j]-rho0[i,j]);
-------------------------------------------------------------------------
Synchronize over all i,j
-------------------------------------------------------------------------
```

(This pseudo-code ignores the consideration at the boundary.) In particular, this algorithmic structure can effectively utilize the parallel computing capabilities of GPUs (and even more so when with the use of ghost cells).

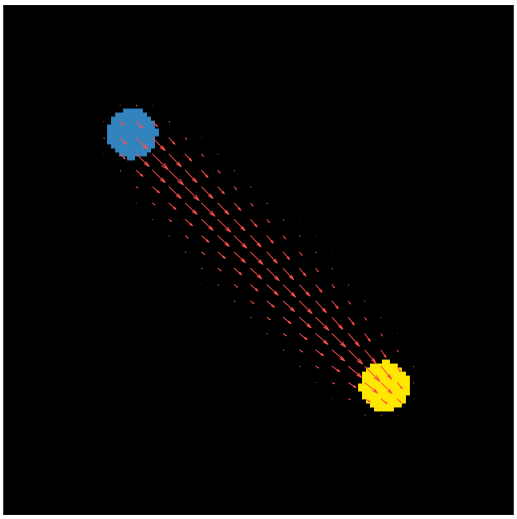
We can use R^k, defined in Section 3.4, as a termination criterion. However, computing R^k can be costly as it requires information from all indices (i, j). So it is best not to compute R^k every iteration.

In choosing the parameters μ and τ Theorem 1 provides an upper bound for the product $\mu\tau$, but does not provide any guidance for their individual values. As they represent the step sizes for the primal and dual variables, quantities of different scales, μ and τ should not be constrained to be equal. Indeed, we have empirically observed that the values of μ and τ must be different by orders of magnitude to get the best convergence rate for both the EMD-L_2 and EMD-L_1 methods and that a poor choice of μ and τ can slow down the rate of convergence significantly. In Section 5, we report the values of μ and τ we used.

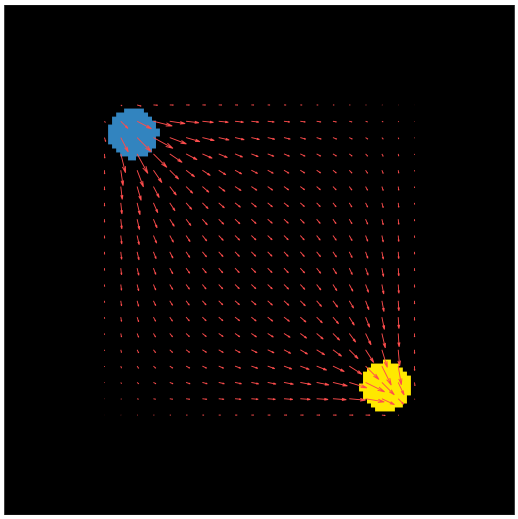
5. Examples

In this section, we demonstrate several numerical results on $\Omega = [-2, 2] \times [-2, 2]$ with an $n \times n$ discretization. The initial values for $\mathbf{m}^0$ and Φ^0 are chosen as all zeros. We implemented the method with CUDA C++ and ran it on the graphics card Nvidia GTX 580 (which costs around $100 as of 2017). We show the flux $\mathbf{m}$ in Figures 1, 2, and 3. We describe the problem description and parameters in the figures' captions. For simplicity, we did not use the termination criterion R^k in these experiments; we simply ran the method up to a fixed iteration count. Rather, we demonstrate the convergence of R^k separately in Figure 4 .

We empirically observe that the methods need roughly $O(n)$ iterations to "converge", where again $n \times n$ is the discretization grid size. This is not surprising as, loosely speaking, information propagates at a rate of one grid point per iteration.

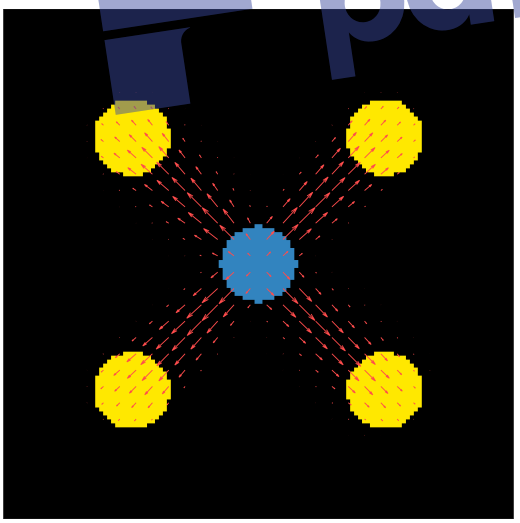


(A) EMD-L_2 solution has value 2.84 and took 1.31s to compute.

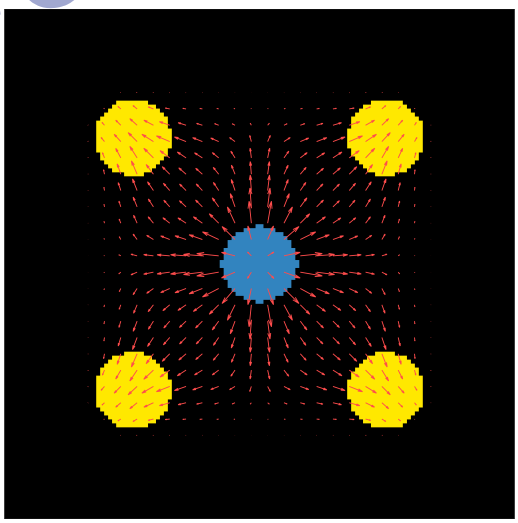


(B) EMD-L_1 solution with $\epsilon = 0.001$ has value 4.00 and took 1.39s to compute.

FIGURE 1. ρ^0 is the blue circle and ρ^1 is the yellow circle. We ran the method with $n = 128$, $\mu = 6 \times 10^{-6}$, $\tau = 6$, and 30,000 iterations.

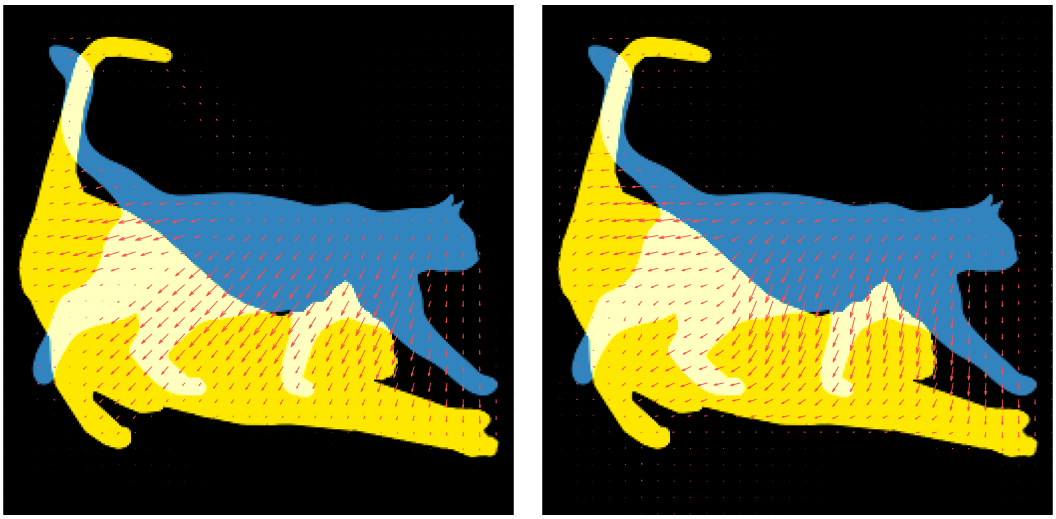


(A) EMD-L_2 solution has value 1.24 and took 1.33s to compute.



(B) EMD-L_1 solution with $\epsilon = 0.001$ has value 1.74 and took 1.38s to compute.

FIGURE 2. ρ^0 is the blue circle and ρ^1 is the yellow circles. We ran the method with $n = 128$, $\mu = 6 \times 10^{-6}$, $\tau = 6$. and 30,000 iterations.



(A) EMD-L_2 solution has value 0.66 and took 16.0s to compute.

(B) EMD-L_1 solution with $\epsilon = 0.001$ has value 0.85 and took 16.0s to compute.

FIGURE 3. ρ^0 is the blue standing cat and ρ^1 is the yellow crouching cat. We ran the method with $n = 256$, $\mu = 3 \times 10^{-6}$, $\tau = 3$, and $100,000$ iterations.

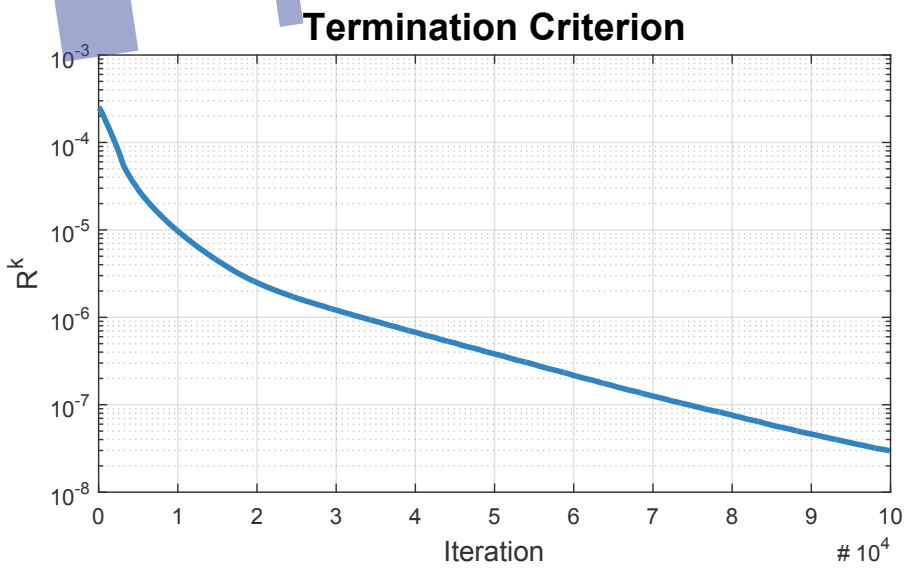
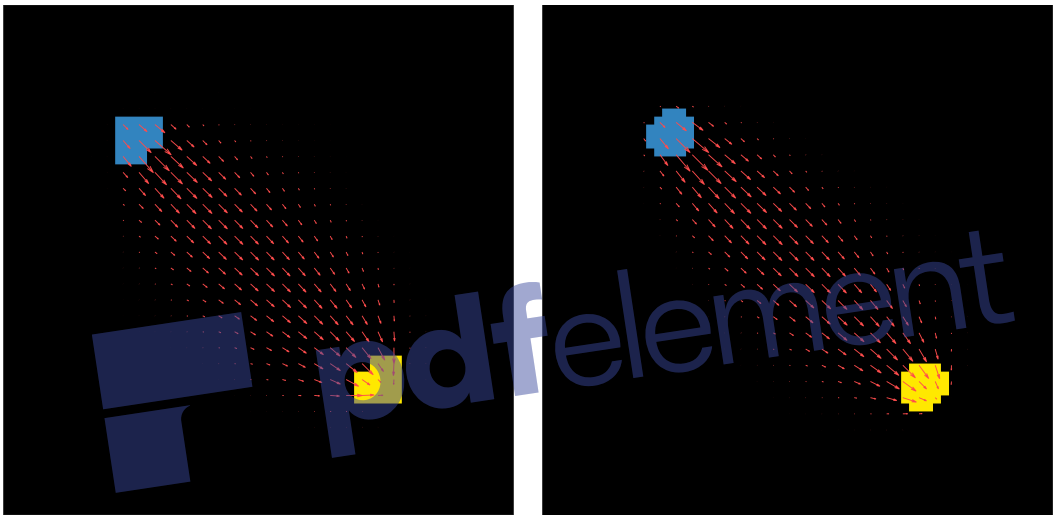


FIGURE 4. Termination criterion R^k for the setup of Figure 3.

However, this observation is somewhat tricky to objectively quantify, as different grid sizes warrant different values of μ and τ. As the definition of the termination criterion R^k

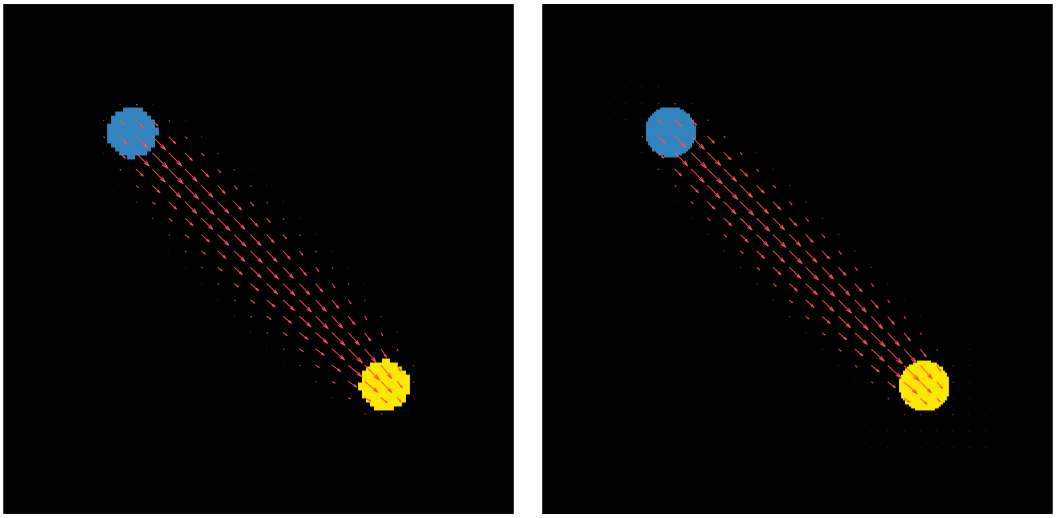
depends on the values of μ and τ, a direct comparison of R^k for setups with different μ and τ provides little information.

So we present a somewhat subjective test to demonstrate this point. The setup is shown in Figure 5. The circles of ρ^0 and ρ^1 are centered at $(-1, 1)$ and $(1, -1)$, respectively, so EMD-L_2 should be roughly $2\sqrt{2} \approx 2.83$. We roughly tuned the parameters μ and τ to get the best performance for each grid size. Finally, we ran the method until the computed EMD-L_2 was close enough to 2.83 and the flux looked good enough. The quantitative results are summarized in Table 1.



(A) $n = 32$, 1000 iterations

(B) $n = 64$, 2000 iterations



(C) $n = 128$, 4000 iterations

(D) $n = 256$, 8000 iterations

FIGURE 5. Testing the number of iterations as a function of grid size.

Grids size	Iteration count	μ	τ	Computed EMD-L_2
32×32	1000	0.0003	3.0	2.876
64×64	2000	0.00007	3.0	2.914
128×128	4000	0.00003	3.0	2.845
256×256	8000	0.000007	3.0	2.752

TABLE 1. Testing the number of iterations as a function of grid size

In Table 2, we compare the wall-clock runtime of the parallel EMD algorithm with other methods. The 4 tested methods are, the presented method run on a GPU (as described at the beginning of this section), the same method implemented in C++ and run serially on an Intel i7 990x CPU, Ling's method [14] run on the same CPU, and Pele's method of [17] run on the same CPU. Pele's method was not able to compute the EMD between inputs larger than 32×32 within a few minutes. We used the 2 cats of Figure 3 (appropriately scaled) for ρ^0 and ρ^1. We also document the number of iterations required until we deemed the method converged.

Grids size	EMD CUDA	EMD CPU	Ling	Pele
32×32	0.012s (1000 iter)	0.08s (1000 iter)	0.007s (600 iter)	$2.74s$
64×64	0.063s (3000 iter)	0.9s (3000 iter)	0.009s (3000 iter)	N/A
128×128	0.336s (10000 iter)	12.9s (10000 iter)	2.3s (30000 iter)	N/A
256×256	6.8s (50000 iter)	245.5s (50000 iter)	80.8s (200000 iter)	N/A

TABLE 2. Runtime of algorithms.

Finally, we mention that the solution to (10) is unique only when $\varepsilon > 0$. We demonstrate this and is in Figure 6. Thus quadratic perturbation is necessary to establish a sense in which the discretized approximations of (10) approximate the true continuous solution as $n \to \infty$.

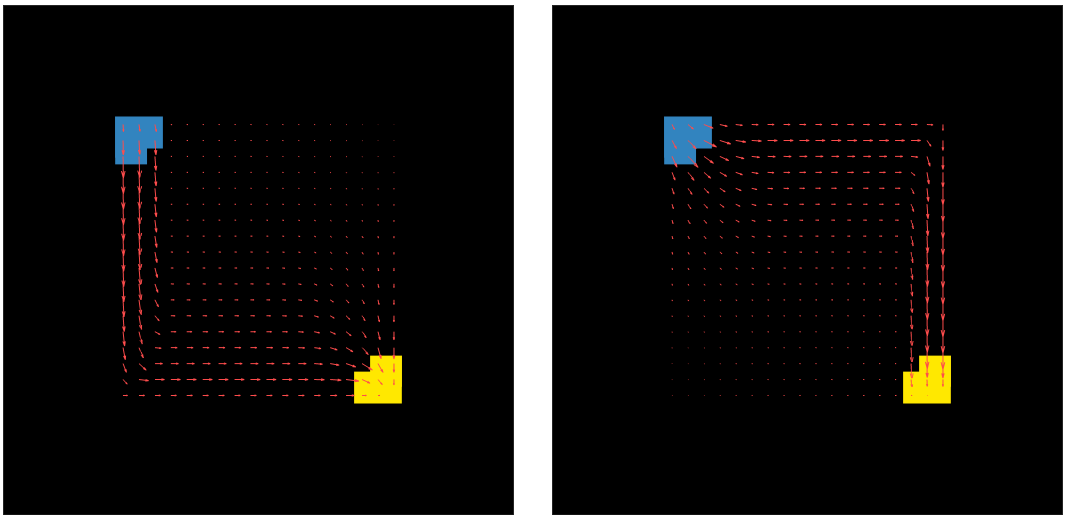


FIGURE 6. Two different solutions for EMD-L_1 when $\varepsilon = 0$.

6. Conclusion

To summarize, we applied a primal-dual algorithm to solve EMD-L_2 and EMD-L_1. The algorithm inherits both key ideas in optimal transport theory and homogeneous degree one regularized optimization problems.

Compared to existing methods, the advantages of proposed algorithm are as follows. First, it leverages the structure of optimal transport, which transfers EMD into a L_1-type minimization, in which the number of variables is much less than the original linear programming problem. Second, it uses simple and parallelizable exact formulas at each iteration (including the shrink operator).

The novel perturbed minimization (10) is computationally useful and deserves attention in future work. In particular, the quadratic regularized term brings some new insights to the original EMD problem. By a direct calculation, one can show that its Euler-Lagrange equation satisfies a pair of partial differential equations:

$$\mathbf{m}(x) = \frac{1}{\epsilon}\left(\nabla\Phi(x) - \frac{\nabla\Phi(x)}{|\nabla\Phi(x)|}\right)$$
$$\frac{1}{\epsilon}\left(\Delta\Phi(x) - \nabla\cdot\frac{\nabla\Phi(x)}{|\nabla\Phi(x)|}\right) = \rho^0(x) - \rho^1(x)\ ,$$

where the second equation holds when $|\nabla\Phi| \geq 1$. Interestingly, the term $\nabla\cdot\frac{\nabla\Phi(x)}{|\nabla\Phi(x)|}$ represents the mean curvature. Another interesting future direction is studying theoretical properties of (10), especially the relationship between minimizers $\mathbf{m}^\epsilon$ and $\mathbf{m}^0$ when ϵ goes to 0.

References

[1] M. Beckmann. A continuous model of transportation, *Econometrica* 20, 643–660, 1952.

[2] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik* 84(3): 375–393, 2000.
[3] Jean-David Benamou and Guillaume Carlier. Augmented Lagrangian methods for transport optimization, mean field games and degenerate elliptic equations. *Journal of Optimization Theory and Applications*, 167(1): 1–26, 2015.
[4] Jean-David Benamou, Guillaume Carlier and Roméo Hatchi. A numerical solution to Monge's problem with a Finsler distance as cost. *M2AN*, 2016.
[5] L.M. Briceño-Arias, D. Kalise and F.J. Silva. Proximal methods for stationary Mean Field Games with local couplings. *arXiv:1608.07701*, 2016.
[6] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 120–145, 2011.
[7] B. Dacorogna and J. Moser. On a partial differential equation involving the Jacobian determinant. *Annales de l'IHP Analyse non linéaire*, 7(1), 1–26, 1990.
[8] Lawrence Evans and Wilfrid Gangbo. Differential equations methods for the Monge-Kantorovich mass transfer problem. *Memoirs of AMS*, no 653, vol. 137, 1999.
[9] Mikhail Feldman and Robert McCann. Monge's transport problem on a Riemannian manifold. *Transactions of the American Mathematical Society*, 354 (4): 1667–1697, 2002.
[10] Tom Goldstein and Stanley Osher. The split Bregman method for L1-regularized problems. *SIAM journal on imaging sciences*, 2(2): 323-343, 2009.
[11] J. Gudmundsson, O. Klein, C. Knauer, and M. Small. Manhattan Networks and Algorithmic Applications for the Earth Movers Distance. *In EWCG*, 2007.
[12] Bingsheng He and Xiaoming Yuan. Convergence Analysis of Primal-Dual Algorithms for a Saddle-Point Problem: From Contraction Perspective, *SIAM Journal on Imaging Sciences*, 5(1), 119–149, 2012.
[13] E. Levina and P. Bickel. The earth mover's distance is the Mallows distance: some insights from statistics *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on* (2):251–256, 2001.
[14] H. Ling and K. Okada. An Efficient Earth Mover's Distance Algorithm for Robust Histogram Comparison. *PAMI*, 2007.
[15] L. Métivier, R. Brossier, Q. Mérigot, E. Oudet and J. Virieux. Measuring the misfit between seismograms using an optimal transport distance: application to full waveform inversion. *Geophysical Journal International*, (205) 1: 345–377, 2016.
[16] Nicolas Papadakis, Gabriel Peyré and Edouard Oudet, Optimal transport with proximal splitting, *SIAM Journal on Imaging Sciences* 7(1): 212–238, SIAM, 2014.
[17] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. *2009 IEEE 12th International Conference on Computer Vision*, 460–467, 2009.
[18] Thomas Pock and Antonin Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization, *2011 International Conference on Computer Vision*, 1762–1769, IEEE.
[19] R. Tyrrell Rockafellar. Conjugate Duality and Optimization, Society for Industrial and Applied Mathematics, 1974.
[20] Yossi Rubner, Carlo Tomasi and Leonidas Guibas. A metric for distributions with applications to image databases. *Computer Vision, 1998. Sixth International Conference on*, 59–66, IEEE, 1998.
[21] Yossi Rubner, Carlo Tomasi and Leonidas Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2): 99–121, 2000.
[22] Leonid Rudin, Stanley Osher and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, (60)1: 259–268, 1992.
[23] Ernest K. Ryu and Stephen Boyd. Primer on Monotone Operator Methods. *Applied and Computational Mathematics*, 15(1):3–43, 2016.
[24] Filippo Santambrogio. Absolute continuity and summability of transport densities: simpler proofs and new estimates. *Calculus of Variations and Partial Differential Equations*, 36 (3): 343–354, 2009,
[25] Sameer Shirdhonkar and David Jacobs. Approximate earth mover's distance in linear time. *Computer Vision and Pattern Recognition IEEE conference*, 2008.
[26] Gilbert Strang. L_1 and L_∞ approximation of vector fields in the plane. *North-Holland Mathematics Studies*, 81, 273–288, 1983.

[27] Justin Solomon, Raif Rustamov, Leonidas Guibas and Adrian Butscher. Earth mover's distances on discrete surfaces. *ACM Transactions on Graphics (TOG)*, 33(4), 2014.
[28] Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.
[29] Wotao Yin, Stanley Osher, Donald Goldfarb and Jerome Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing, *SIAM Journal on Imaging sciences*, 1(1): 143–168, 2008.

E-mail address: wcli@math.ucla.edu

E-mail address: eryu@math.ucla.edu

E-mail address: sjo@math.ucla.edu

E-mail address: wotaoyin@math.ucla.edu

E-mail address: gangbo@math.ucla.edu

Department of Mathematics, University of California, Los Angeles.