Image Classification Using Wasserstein Distance from Monge-Kantorovich Solvers

Ruiqi Wang, Mingzhen Li, Anthony Hong, Jingyuan Zhu

December 14, 2023

Abstract

In this course project, we explore the application of Optimal Transportation theory in the form of Nonlinear Monge-Kantorovich Problem, and its application in grayscale digit classification on the MNIST dataset. We studied the Monge-Kantorovich Problem with a quadratic cost function and wrote the problem in Primal and Dual problem forms. Two solutions for both forms are studied and implemented. We further investigate the role of Wasserstein distance and other distance metrics in a distance-based classification method, K-nearest neighbors (KNN), and compare their effect on classification accuracy.

1 Introduction

In our project, we explore image classification, a key aspect of machine learning, using the standard benchmark, MNIST dataset. The primary challenge lies in accurately identifying and categorizing images based on their inherent visual features. We investigate Optimal Transportation via the Nonlinear Monge-Kantorovich Problem, which transits image classification into a cost minimization to measure image differences and similarities. This approach of similarity measurement originates from the mathematical concept of efficiently transferring mass between distributions.

Figure 1 visually interprets the Optimal Transportation problem. Consider a pile of dust (on the left of the figure), denoted by μ, that needs to be moved from its original domain X, e.g., on the grass, to fill the hole in a sandbox, denoted by ν (on the right of the figure), within a new domain Y, e.g., the sandbox. The task is to transport each grain of dust from a point x in domain X to a point y in domain Y, incurring a cost $c(x, y)$ for each grain of movement. The objective is to determine the most cost-efficient strategy to relocate the entire distribution μ into the sandbox ν, such that the final arrangement of dust replicates the desired distribution, i.e., to fill the holes in the sandbox. This encapsulates the essence of the Monge-Kantorovich Problem, where one seeks the least costly way to transform one distribution into another.

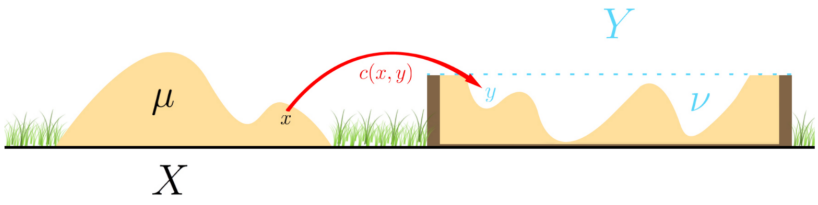

Figure 1: Visual representation of the Optimal Transportation problem using the Monge-Kantorovich theory.

The core of image classification and similarity calculation with the Monge-Kantorovich Problem is treating the images as probability distributions, and the similarity between images is essentially the cost of transportation from one image to another. In the specific case of grayscale digit images in the MNIST dataset, we treat the grayscale image inputs as probability distributions. For simplicity, we assume X and Y to be 2D spaces, and the pixels are restrained in squares for simplicity. We convert images into 2D probability distributions that reflect images' pixel intensities (after normalization) through

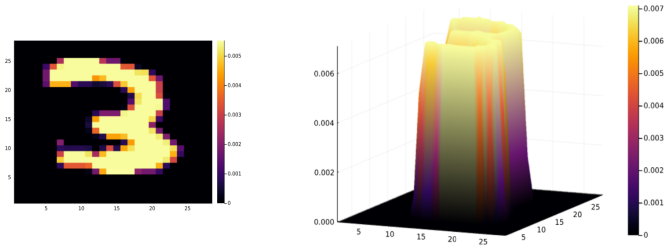

Figure 2: Converting normal MNIST digits into a 2D probabiliy distribution. The original MNIST digit after normalization is depicted on the left and the visualization of the converted 2D distribution is shown on the right.

normalization and other tricks, as in Figure 2. In this sense, the solution to the Monge-Kantorovich Problem is the Wasserstein distance between the image pairs. The Wasserstein metric, especially in its application to the Monge-Kantorovich Problem, provides a holistic approach to measuring distances between probability distributions. As a result, we conducted a detailed examination of Wasserstein distance within the context of distance-based image classification with the common KNN algorithm.

In our project's problem formulation, we focus on the formulation of the Monge-Kantorovich Problem with a quadratic cost function, a common choice that penalize more on longer transportations so similar images would have even shorter distances. This allows us to formulate the image classification problem into a non-linear problem. In this project, we solved the problem in its primal and dual forms, establishing a robust theoretical foundation from two mutually supportive directions. Additionally, we will compare the results of KNN using other traditional distance metrics to assess the impact of these different metrics on classification accuracy.

This paper is structured as follows: In section 2, we provide formal mathematical formulations in both the primal and dual forms. To solve the problem with images as inputs, we provide corresponding solutions in section 3. The mathematical foundation is then followed by a complexity analysis for computational complexity analysis in section 4. Then in section 5 we show how the digit classification problem can be formulated as a 2D Monge-Kantorovich Problem by treating the grayscale images as probability distributions and in section 6 we show the experimental results of image classification accuracy using the proposed method. Then we conclude the paper with a discussion.

In summary, our project has the following contribution: we creatively use the Optimal Transportation and the Nonlinear Monge-Kantorovich Problem to tackle image classification. By integrating recent advances in mathematics with coding implementations, we enhanced the accuracy of digit classification in the MNIST dataset without neural network training with a large amount of training samples. This project helps us better understand optimal transportation and gain insights into the power of non-linear optimizations in various real-world applications.

2 Monge-Kantorovich Problem and Wasserstein Distance

Before giving a classic formulation of Monge and Kantorovich's problems, we first define some basic concepts in line with [1] sections 5.2 and 6.1.

If X, Y are separable metric spaces, $\mu \in \mathscr{P}(X)$, the set of all probability measures on $(X, \mathscr{B}(X))$, and $T : X \to Y$ is a Borel-measurable map, we define the **push-forward** of μ through T by

$$T_{\#}\mu(B) := \mu\left(T^{-1}(B)\right) \quad \forall B \in \mathscr{B}(Y). \tag{1}$$

The measure $T_{\#}\mu$ belongs to $\mathscr{P}(Y)$ and is also called the **image measure**. An equivalent definition is that for every bounded (or $T_{\#}\mu$-integrable) Borel function $f : Y \to \mathbb{R}$,

$$\int_X f(T(x))d\mu(x) = \int_Y f(y)dT_{\#}\mu(y) \tag{2}$$

A **coupling**, or **transport plan**, of two probability measures μ and ν on the measurable spaces $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ respectively is any probability measure γ on the product measurable space $(X \times Y, \mathcal{X} \otimes \mathcal{Y})$ (where $\mathcal{X} \otimes \mathcal{Y}$ is the smallest σ-algebra containing $\mathcal{X} \times \mathcal{Y}$) whose marginals are μ and ν, i.e.,

$$\gamma(A \times Y) = \mu(A) \quad \forall A \in \mathcal{X}$$
$$\gamma(X \times B) = \nu(B) \quad \forall B \in \mathcal{Y}$$

Usually, $\mathcal{X} = \mathscr{B}(X)$ and $\mathcal{Y} = \mathscr{B}(Y)$. Note that if we let $\pi^i, i = 1, 2$, be the projections from $X \times Y$ to index spaces, we have $A \times Y = (\pi^1)^{-1}(A)$ and $X \times B = (\pi^2)^{-1}(B)$. Hence, the above **marginality condition** is equivalent to

$$\mu = \gamma \circ (\pi^1)^{-1} = \pi^1_{\#}\gamma$$
$$\nu = \gamma \circ (\pi^2)^{-1} = \pi^2_{\#}\gamma$$

We denote all such transport plans of μ and ν as $\Gamma(\mu, \nu)$:

$$\Gamma(\mu, \nu) := \left\{\gamma \in \mathscr{P}(X \times Y) : \pi^1_{\#}\gamma = \mu, \pi^2_{\#}\gamma = \nu\right\}.$$

Notice also that $\Gamma(\mu, \nu) = \{\mu \times \nu\}$ if either μ or ν is a Dirac measure.

2.1 Primal and Dual in the Monge-Kantorovich Problem

Now, let X, Y be Polish spaces (i.e., separable completely metrizable topological spaces) and $c : X \times Y \to [0, +\infty]$ be a Borel cost function (we note that Radon spaces, spaces such that Borel probability measures are tight, are considered in [1] but are one of the generalizations of the Polish spaces, spaces sufficiently nice to cover our image classification application). Given $\mu \in \mathscr{P}(X), \nu \in \mathscr{P}(Y)$ the optimal transport problem, in Kantorovich's formulation, is given by

$$(K): \quad \min\left\{\mathcal{K}[\gamma] = \int_{X\times Y} c(x, y)d\gamma(x, y) : \gamma \in \Gamma(\mu, \nu)\right\} \tag{3}$$

(K) is posed as a relaxation of (M) defined below, which may have no transport at all not to mention an optimal one. That is, in language of optimization theory, the feasible set by above constraint can be empty. Fortunately, (K) solves this issue: (1) there is an least one transport in $\Gamma(\mu, \nu)$, which is just the product measure $\mu \times \nu$; (2) if c is bounded and continuous, and if μ has no atom, then the "min" in (3) will be equal to "inf" in (4):

$$(M): \quad \inf\left\{\mathcal{M}[T] = \int_X c(x, T(x))d\mu(x) : T \in \mathcal{T}(\mu, \nu)\right\}. \tag{4}$$

where $\mathcal{T}(\mu, \nu) = \{\text{measurable } T : X \to Y | T_{\#} = \nu\}$

To solve (K), one often considers its dual form:

$$(D): \quad \max\left\{\mathcal{D}[u, v] = \int_X u(x)\mathrm{d}\mu(x) + \int_Y v(y)\mathrm{d}\nu(y) : (u, v) \in \Phi_c(u, v)\right\} \tag{5}$$

where $\Phi_c(u, v) = \{(u, v) \in \mathcal{L}^1(\mu) \times \mathcal{L}^1(\nu) : u(x) + v(y) \leq c(x, y)\}$

In particular, the Monge-Kantorovich problem in discrete setting is given by the Birkhoff theorem, where $\Gamma(\mu, \nu)$ is a convex set of all doubly stochastic $N \times N$ matrices (see [1] Theorem 6.0.1 for proof and also Theorem 6.0.2 for another special case where $X = Y = \mathbb{R}$).

2.2 Wasserstein Distance

(M),(K),(D) above are called *problems* in the sense that their goals are to find the argument that achieves the optimum value, while Wasserstein distance is the square root of the optimum value.

Let X, Y, μ, ν be given as in the last subsection. The **Wasserstein distance** between μ and ν associated with (M) with quadratic c is

$$W(\mu,\nu) = \left(\int_X \frac{1}{2}|x - T^*(x)|^2 f(x)dx\right)^{1/2} \tag{6}$$

where T^* is the optimal transportation minimizing (M).

Remark 2.1. One can consult the reference [1], in particular chapter 7, to see that the above definition indeed gives a distance on the space of probability measures $\mathcal{P}(X)$: symmetry is obvious; positivity is checked because $W(\mu,\nu) = 0$ implies $\exists\gamma \in \Gamma(\mu,\nu)$ s.t. $\int d(x,y)d\gamma(x,y) = 0$ and thus $x = y$ γ-a.e., implying that

$$\pi(\text{ Diagonal }) > 0, \pi\left(X^2 - \text{ Diagonal }\right) = 0 \Rightarrow \gamma = \mathrm{Dup}_{\#}\,\mu$$

where Dup is the duplicate $x \mapsto (x,x)$, which by composition rule (1) gives $\nu = \pi^2_{\#}\gamma = \mathrm{id}_{\#}\mu = \mu$; triangle inequality requires a bit more terminology, namely the composition of transportation plans (see [1] Remark 5.3.3). However, the triangle inequality is intuitive in our mass-moving narration. Plan A moves masses from place (dist.) μ to place (dist.) ν and plan B from ν to λ, both minimizing the costs on their own, while the optimal plan C minimizes the cost of moving masses from μ to final destination λ, without first going to some intermediate place (dist.) ν.

2.3 Theoretical Solution of Monge's Problem

In our paper, we consider the case that $X = Y = \mathbb{R}^d$ and $\mu, \nu \in \mathscr{P}\left(\mathbb{R}^d\right)$ with the quadratic cost function

$$c(x, T(x)) = \frac{1}{2}|x - T(x)|^2$$

If we write μ and ν in their Radon-Nikodym forms, we have

$$\mathrm{d}(x)\mu = f(x)\mathrm{d}x, \quad \mathrm{d}(y)\nu = g(y)\mathrm{d}y$$

The primal problem of (M) with a quadratic cost function is then formulated in the following way

$$\begin{aligned} &\inf \int_X \frac{1}{2}|x - T(x)|^2 \mathrm{d}\mu(x) \\ &\text{s.t. } T_{\#}\mu = \nu \end{aligned} \tag{7}$$

The following equivalences will serve as a basis for our discussion.

$$T_{\#}(\mu) = \nu \text{ (Constraint in (M))} \tag{8}$$

$$\Leftrightarrow \int_X h(T(x))f(x)\mathrm{d}x = \int_Y h(y)g(y)\mathrm{d}y \quad \forall h \in \mathscr{B}^d(X) \tag{9}$$

$$\Leftrightarrow f(x) = g(T(x))\det(\mathrm{D}T(x)) \text{ (Monge-Ampere Equation)} \tag{10}$$

The first equivalence is due to (2) and we use $h \in \mathscr{B}^d(X)$ to say that h is Borel measurable and bounded; the third equivalence is due to change of variable formula, where D is the Jacobian of the mapping T.

Now the celebrated result of Brenier's is given below:

Theorem 2.2. [Brenier Theorem]* Let $X = Y = \mathbb{R}^d$ and assume that μ, ν both have finite second moment such that μ does not give mass to small sets (those ones with Hausdorff dimension are at most $d-1$). The cost function is $c(x,y) := \frac{1}{2}|x-y|^2$. Then

1. For each γ^* minimizing $\mathcal{K}[\gamma]$ over $\Gamma(\mu,\nu)$, there is T minimizing $\mathcal{M}[T]$ over $\mathcal{T}$ such that $\gamma^* = (\mathrm{Id}, T)_\sharp\mu$.
2. For each $T \in \mathcal{T}$, i.e., a solution of Monge-Ampere eq., T minimizes $\mathcal{M}[T]$ over $\mathcal{T}$ if and only if $T = \nabla\varphi$ for some proper convex l.s.c. $\varphi \in \mathcal{L}^1(\mu)$.

*Original formulation in [3], current version formulated by this post.

3. There is a unique (up to μ-a.e.) minimizer T of $\mathcal{M}[T]$ over $\mathcal{T}$.

Note that (D) with quadratic cost function $c(x,y) = \frac{1}{2}|x-y|^2$, by a change of variable

$$\begin{cases} \phi(x) := \frac{1}{2}|x|^2 - u(x) & (x \in X) \\ \psi(y) := \frac{1}{2}|y|^2 - v(y) & (y \in Y). \end{cases}$$

is equivalent to

$$(L): \quad \min_{(\phi,\psi)} \left\{ \mathcal{L}[\phi,\psi] = \int_X \phi(x)\mathrm{d}\mu(x) + \int_Y \psi(y)\mathrm{d}\nu(y) : \phi(x) + \psi(y) \geq x \cdot y \right\} \tag{11}$$

Lemma 3.1 and Theorem 3.1 of [4] shows that there exists solution pair (ϕ^*, ψ^*) solving the dual (11). and $s^* = \mathrm{D}\phi^* = \nabla\phi^*$ recovers the solution of the primal (7):

Theorem 2.3 (Evans's)**.** There exist a unique minimizing pair (ϕ, ψ) solving (L) 11, and (ϕ, ψ) are convex conjugates of each other, that is

$$\begin{cases} \psi(x) = \phi^*(x) := \max_{y \in Y} (x \cdot y - \psi(y)) & (x \in X) \\ \phi(y) = \psi^*(y) := \max_{x \in X} (x \cdot y - \phi(x)) & (y \in Y). \end{cases}$$

If we define $s = \mathrm{D}\phi$, then

1. $\mathbf{s} : X \to Y$ is essentially one-to-one and onto (this allows us to do image classification)
2. $\int_X h(\mathbf{s}^*(x))\, d\mu(x) = \int_Y h(y) d\nu(y)$ for each $h \in C(Y)$.
3. S solves (M).

3 Algorithmic Solution of Monge's Problem

Now, we have two theoretical (and algorithmic) approaches to solve the Monge-Kantorovich problem. and thus to find Wasserstein distance, W:

(a) Paper [7] (where Proposition 2.1 in this paper is the same as Evans's 2.3) gives an explicit computation of the differential of the functional $\mathcal{L}[\phi,\psi]$ in its Theorem 3.1. Theorem 2.3 tells us that $\mathcal{L}$ has unique minimizing pair of the form (ϕ, ϕ^*), so plugging it into the functional will get

$$\mathcal{L}[\phi] = \mathcal{L}[\phi, \phi^*] = \int_X \phi \mathrm{d}\mu + \int_Y \phi^* \mathrm{d}\nu$$

The differential of $\mathcal{L}$ with respect to ϕ is

$$\mathcal{L}'[\phi] = f - (g \circ \nabla\phi^{**}) \det\left(\mathrm{D}^2\phi^{**}\right)$$

Naturally, gradient descent with the following update rule ([7] eq. (15), (24)) can be considered:

$$\phi_{n+1} = \phi_n - \alpha_n \mathcal{L}'[\phi_n] = \phi_n - \alpha_n \left(f - (g \circ \phi_n) \det\left(\mathrm{D}^2\phi_n\right)\right) \tag{12}$$

After the minimizer ϕ is found, theorem 2.3 claims that $s = \mathrm{D}\phi$ is the solution of (M) and thus defines (6).

(b) Paper [6] assumes $T = \mathrm{D}(\phi)$ where $\phi = \frac{1}{2}|x|^2 - u$ for some u (thus $T = x - \mathrm{D}u$) and plug into the Monge-Ampere equation

$$f(x) = g(T(x))\det(\mathrm{D}T(x))$$

to get

$$\det\left(I - \mathrm{D}^2 u\right) g(x - \mathrm{D}u) = f(x) \tag{13}$$

The paper then uses linearization and discretization of pde to find u that solves the two-dimension form of the above equation (14). Finding such u will give rise to $T = x - \mathrm{D}u$, which solves the Monge-Ampere equation (9) and thus satisfies constraint (7), i.e., $T \in \mathcal{T}$. Therefore, by Brenier's Theorem 2.2 second claim, T solves (M) and defines 6.

3.1 Gradient Descent Solution

In our algorithm, we closely follow the methodology outlined for addressing the Monge-Kantorovich problem. This theoretical basis is implemented computationally, incorporating heuristic choices for initial conditions of the potential function ϕ and gradient descent step sizes (α_n). These heuristic parameters are tailored to enhance the algorithm's efficiency and convergence. The algorithm employs iterative gradient descent updates of ϕ, aligning with the Monge-Kantorovich framework until convergence is achieved within a specified tolerance. This approach blends theoretical foundations with practical computational strategies.

```
import numpy as np

def compute_L_prime(phi, f, g):
    return f - (g * np.abs(np.linalg.det(np.gradient(phi))) * np.gradient(g))

def gradient_descent_update(phi, alpha_n, f, g):
    return phi - alpha_n * compute_L_prime(phi, f, g)

def compute_wasserstein_distance(phi, mu, X):
    T = np.gradient(phi)
    return np.sqrt(np.trapz(0.5 * np.linalg.norm(X - T, axis=1)**2 * mu, X))

# Initialization
phi = np.zeros_like(X)
alpha_n = 0.01
max_iterations = 1000
convergence_threshold = 1e-6

# Gradient Descent Loop
for iteration in range(max_iterations):
    phi_old = phi.copy()
    phi = gradient_descent_update(phi, alpha_n, mu, nu)
    if np.linalg.norm(phi - phi_old) < convergence_threshold:
        break

# Output
print("Wasserstein Distance:", compute_wasserstein_distance(phi, mu, X))
```

3.2 Monge-Ampère Equation Solution

To employ Newton numerical solution for an approximation to the Wasserstein distance between two images, two main issues must be resolved. The first issue is that the PDE formulation assumes smooth, continuous density functions, but the image pixels are discrete. This issue is addressed by discretization. The second issue is that PED formulation assumes strongly positive density, but it is common for an image representation to have pixels of intensity zero. To resolve this issue, paper [6] add a small constant to the image before normlaization. In addition, two assumptions were made before solving equation (14). First, the images to be compared are both the same size. Second, each image is defined on the unit square, so $X = Y = [0, 1]^2$. The unit square is taken to be oriented with positive derivative in the x_1 and x_2 directions following the x_1 and x_2 axis respectively.

Then, we can solve the two-dimensional Monge-Ampère equation (14)

$$\det\left(I - \mathrm{D}^2 u\right) g(x - \mathrm{D}u) = f(x) \tag{14}$$

where both x and u are in two dimensions. As we want to map the edges of each image to the other as well as the corners, the equation is subject to homogeneous Neumann conditions at the boundary, i.e., $\frac{\partial X}{\partial \mathbf{n}} = 0$ where $\mathbf{n}$ is normal to the boundary.

3.2.1 Linearization

To solve the equation, we first perform linearization by the transformation $u \approx u + \epsilon w$ for small ϵ and some initial guess u. Then the equation (14) becomes

$$det(I - D^2(u + \epsilon w))g(x - D(u + \epsilon w)) = f(x) \tag{15}$$

Let $\alpha = 1 - u_{x_2x_2} - u_{x_1x_2}u_{x_2x_2} - u_{x_1x_2}^2$; $\beta = u_{x_2x_2} - 1$; $\gamma = u_{x_1x_1} - 1$; $\delta = -2u_{x_1x_2}$

Then, with the new notation, expanding equation (15) while neglecting terms of order ϵ^2 or higher, we get

$$det(I - D^2(u + \epsilon w)) = (\alpha + \beta\epsilon w_{x_1x_1} + \gamma\epsilon w_{x_2x_2} + \delta\epsilon w_{x_1x_2}). \tag{16}$$

Taking the Taylor expansion of $g(x)$ around $x - u_x$ and denoting $G = g(x_1 - u_{x_1}, x_2 - u_{x_2})$ for conciseness, then we get

$$g(x_1 - u_{x_1} - \epsilon w_{x_1}, x_2 - u_{x_2} - \epsilon w_{x_2} \approx G - G_{y_1}\epsilon w_{x_1} - G_{y_2}\epsilon w_{x_2}. \tag{17}$$

Then, according to equation (15), combining (16) with (17) gives the linearized equation

$$(\alpha + \beta\epsilon w_{x_1x_1} + \gamma\epsilon w_{x_2x_2} + \delta\epsilon w_{x_1x_2}) \cdot (G - G_{y_1}\epsilon w_{x_1} - G_{y_2}\epsilon w_{x_2}) = f(x_1, x_2). \tag{18}$$

Multiplying out the above equation and neglecting terms of order ϵ^2 or higher, we get

$$G(\beta\epsilon w_{x_1x_1} + \gamma\epsilon w_{x_2x_2} + \delta\epsilon w_{x_1x_2}) - \alpha(G_{y_1}\epsilon w_{x_1} + G_{y_2}\epsilon w_{x_2}) = f(x_1, x_2) - \alpha G. \tag{19}$$

This is a second order linear partial differential equation in ϵw. For an initial guess of the solution u, we iteratively solve for ϵw, which is used to update the solution $u_{n+1} = u_n + \epsilon w$.

3.2.2 Discretization

To solve the linearized equation (19) numerically, paper [6] consider standard central differences on a uniformly discretized grid. In the case of images, an image is considered as a discretization of a surface on $[0, 1]^2$. When evaluating G and its derivatives in equation (19) at point $(x_1 - u_{x_1}, x_2 - u_{x_2})$, the will generally not lie on a known pixel point. To resolve this issue, the paper perform an interpolation on the pixels to construct continuous function from the discrete representation of the image.

For a general function ψ, we denote its position on the grid by $\psi_{i,j}$ for $i = 1, 2, .., m$ and $j = 1, 2, .., n$, where m and n are the number of grid points in the x_1 and x_2 axis respectively. Paper [6] adopted the following central differences with step sizes h and k on the uniformly discretized grid to solve equation (19) numerically:

$$\begin{aligned}
\psi_{x_1} &= \frac{\psi_{i+1,j} - \psi_{i-1,j}}{2h},\ \psi_{x_2} = \frac{\psi_{i,j+1} - \psi_{i,j-1}}{2k};\\
\psi_{x_1x_1} &= \frac{\psi_{i-1,j} - 2\psi_{i,j} + \psi_{i+1,j}}{h^2},\ \psi_{x_2x_2} = \frac{\psi_{i,j-1} - 2\psi_{i,j} + \psi_{i,j+1}}{k^2};\\
\psi_{x_1x_2} &= \frac{\psi_{i+1,j+1} - \psi_{i+1,j-1} - \psi_{i-1,j+1} + \psi_{i-1,j-1}}{4hk}.
\end{aligned} \tag{20}$$

Following the above central differences to discretize ϵw, the equation for an interior point in the grid in given by

$$\begin{aligned}
&- 2G\left(\frac{\beta}{h^2} + \frac{\gamma}{k^2}\right)\epsilon w_{i,j} + \left(\frac{\beta G}{h^2} + \frac{\alpha G_{y_1}}{2h}\right)\epsilon w_{i,j-1}\\
&+ \left(\frac{\beta G}{h^2} - \frac{\alpha G_{y_1}}{2h}\right)\epsilon w_{i,j+1} + \left(\frac{\gamma G}{k^2} + \frac{\alpha G_{y_2}}{2k}\right)\epsilon w_{i-1,j}\\
&+ \left(\frac{\gamma G}{k^2} - \frac{\alpha G_{y_2}}{2k}\right)\epsilon w_{i+1,j} + \frac{\delta G}{4hk}(\epsilon w_{i+1,j+1} - \epsilon w_{i+1,j-1} - \epsilon w_{i-1,j+1} - \epsilon w_{i-1,j-1})\\
&= f_{i,j} - \alpha G
\end{aligned} \tag{21}$$

where each of the coefficients are known values from the initial guess u.

Write the equation in matrix multiplication form, we get

$$A\epsilon\mathbf{w} = b$$

where A is an $n \times n$ coefficient matrix from the left-hand side of equation (21), and b is the right-hand side of equation (21). Since imposing homogeneous Neumann conditions results in a rank of $n-1$ for the coefficient matrix, paper [6] impose an extra condition that the solution is orthogonal to the kernel, which gives us

$$\begin{bmatrix} A & e \\ e' & 0 \end{bmatrix} \begin{bmatrix} \epsilon w \\ \lambda \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}.$$

This system of equations can be solved by a LP solver, and the solution u can be updated accordingly. After the solution u subject to some threshold of tolerance is found, we get the optimal transport of the form $T = x - Du$ according to Brenier's Theorem. Then, we get the Wasserstein distance

$$W = \left(\int_X \frac{1}{2} |x - T(x)|^2 f(x) dx \right)^{1/2}$$

4 Computational Complexity

We conducted computational complexity analysis for both solutions. The details are as following:

- **Numerical Approach to the Monge-Ampère Equation:** We analyze the complexity of this method in a theoretical manner, following the analysis in [6]. The computational complexity is principally influenced by three factors: setting up the sparse matrix, solving the linear system, and the iterative process to achieve the desired accuracy. The most complex element lies in solving the large-scale linear system. Efficient utilization of the sparse nature of the coefficient matrix by solvers typically leads to a complexity nearing $O(N \log(N))$, with N representing the pixel count. After the PDEs are linearized, linear programming techniques, such as the simplex algorithm, are applied with a theoretical worst-case complexity that is polynomial in nature. However, in practical scenarios, they tend to execute much quicker. When considering the combined complexities of all elements in the solution and other additional overhead from initialization and iterative refinement, the total computational load of this approach is anticipated to fall in the range of $O(N)$ to $O(N^2)$.
- **Gradient Descent Solution:** In light of the time constraints of our project, conducting a detailed complexity analysis of the gradient descent method, especially concerning the approximation error ϵ, is unfeasible. As such, we defer to the findings of [2] for an in-depth computation. The work shows their gradient based approach with a complexity of $O(N^{2.5}\sqrt{\log N}/\epsilon)$ for an error threshold of ϵ. Other methods [8, 9], also reported a general complexity between $O(N^2)$ and $O(N^3)$ when simplifying the error ϵ as a constant. As a result, we would expect the Gradient Descent approach to be much more complex than the Numerical approach. This is also reflected in the actual experiments.

5 Application to Image Classification

The following experiment investigates the effectiveness of employing Wasserstein distance metric in image classification. In this context, for each pair of given image f and hypothesis image g (in the hypothesis set of all images), we compute a Wasserstein distance in both solvers and compare their performances. In addition, we include two widely used distance metrics to perform the same prediction task with sharing KNN algorithm to evaluate the effectiveness of using the Wasserstein distance metric in KNN for the prediction task on MNIST dataset.

5.1 Dataset

The "Modified National Institute of Standards and Technology database" (also widely known as the MNIST dataset) is a widely used collection of handwritten digit images, which has become a benchmark in the field of machine learning and computer vision. It contains 60,000 training images and 10,000 testing images of handwritten digits. Each image is a grayscale 28x28 pixel image. Figure 4 shows examples of digits in the MNIST dataset. The MNIST dataset has been extensively utilized for

developing and testing various image-processing techniques and classification algorithms and plays a crucial role in the development and evaluation of machine learning models.

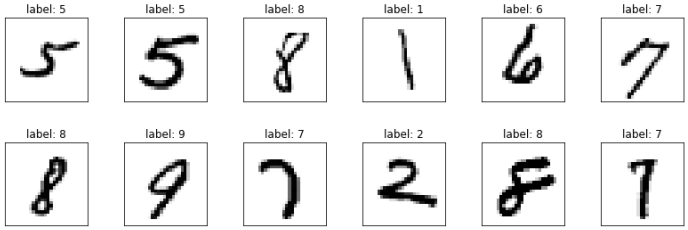

Figure 3: MNIST digit examples.

The Monge-Kantorovich Problem is applicable in applications where the data can be treated as distributions (in MNIST's case, pixel intensities across 784 entries). The benefit of using the MNIST is that the grayscaled images provides an excellent opportunity to treat the inputs as distributions and allows us to demonstrate how the Monge-Kantorovich approach can effectively classify images based on the underlying distribution of their pixel intensities.

5.2 Experimental Setup

In our experiment, given the constrained computational resources, we opted to work with a subset of the MNIST dataset. As discussed in [6] the Wasserstein Distance out performs other distance metrics when the size of the training set is limited. As a reuslt, we randomly sampled from the MNIST dataset training set of various sizes comprising 1, 5, 10, 20, and 30 images for each digit (so the number of training images will be 10x) and test sets are constructed in the same way as the training set. Both the training and test sets underwent normalization procedures to ensure that the numerical integration of the image pixels equated to 1, as required by a probability distribution.

We employed the KNN method to classify each input image by comparing the distance between the input and all the training images. We use a fixed value of $K = 1$ for the K parameter of the KNN method. Three distance metrics are used for distance estimation:

1. **Wasserstein Distance**: We computed the Wasserstein distance (6) as a measure of similarity between images using the MK pb. solver, i.e., $W(\mu, \nu) = \left(\int_X \frac{1}{2}|x - T^*(x)|^2 f(x)dx\right)^{1/2}$. This method takes into account the distribution of pixel intensities in the images, especially spatial relationships and their distribution across the entire image, rather than the pixel intensities alone in the conventional distance metrics, such as the Euclidean Distance.

2. **Euclidean Distance**: Euclidean distance is a direct measure of the straight line distance between two points. It is calculated using the square root of the sum of the squared differences between the corresponding coordinates or values of the points, $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$.. This metric is commonly used in computer vision for measuring the similarity between images. It is relevant to the mean square error.

3. **Pearson's Linear Correlation Coefficient** Pearson's Linear Correlation Coefficient, r, is a statistical measure that shows the strength and direction of a linear relationship between two variables, $d(\mathbf{x}, \mathbf{y}) = r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (x_i - \overline{x})^2}}$. It is widely used in various fields to understand and predict relationships between variables. It ranges from -1 (strong negative correlation) to +1 (strong positive correlation), with 0 indicating no linear correlation. This distance metric can be calculated by dividing the covariance of the variables by the product of their standard deviations if we consider two images to be two variables represented by matrices.

As we aimed to assess the performance of each metric with varying training set sizes, ranging from small to large, as indicated in the figure 4. To achieve this, we firtly applied the KNN algorithm with the smallest set, consisting of only one instance for each digit. Subsequently, we iteratively conducted the same test for the three specified distances, progressively augmenting the number of occurrences for each digit in the training set. For instance, in the concluding test, the training set comprised 30

instances of the digit "one," 30 instances of the digit "two," and so forth, resulting in a total training set size of 300. We conducted this experiment across all differently constructed training sets to account for dataset variability. The metric used to evaluate each method is the image classification accuracy.

Notice that although we have introduced two different kinds of algorithm for computing Wasserstein distance, the numerical way is the only method we have adopted for calculation. The reason is that the GD approach is extremely computationally inefficient, as discussed in Section 4, which results in impossibility of the required computational task on a 2D space. For instance, given the training and testing dataset of 10 samples, the GD approach requires complete Gradient Descent procedure for each test and sample in order to generate a value for wasserstein distance, leading to performing 100 complete gradient descent processes. Such heavy computation costs prohibit it from any practical use. Therefore, we only utilize numerical approach for the sake of stability and efficiency.

6 Experimental Results

In this section, we present the evaluation results of our image classification experiment on the MNIST dataset using the KNN method with three different distance metrics: the Monge-Kantorovich problem solved Wasserstein distance, Pearson's Linear Correlation Coefficient and the Euclidean distance.

6.1 Accuracy

We evaluated the performance of our KNN classifier using both distance metrics and calculated the accuracy of the classification on the test dataset.

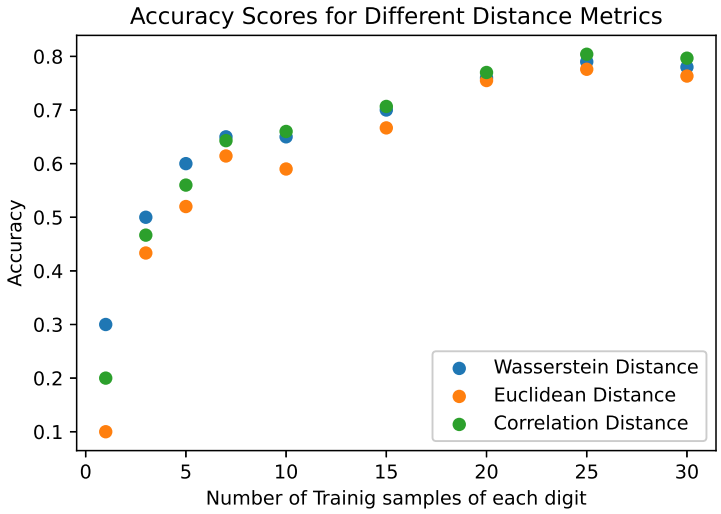

Figure 4: Accuracy for KNN using different distance metrics.

Table 1: Image Classification Accuracy Scores for Different Distance Metrics. The bold numbers represent the best performance among all three methods. The underlined numbers represent the second-to-best accuracy.

	Number of Training Samples per Digit				
Distance Metric	1	5	10	20	30
Euclidean	0.1	0.52	0.59	0.76	0.76
Correlation	<u>0.2</u>	<u>0.56</u>	**0.66**	**0.77**	**0.8**
Wasserstein	**0.3**	**0.6**	<u>0.65</u>	<u>0.76</u>	<u>0.78</u>

As shown in Table 1, the KNN classifier using the Wasserstein distance metric achieved high accuracy when the number of training samples was low. For instance, the Wasserstein distance metric has a

score of 30%, outperforming the KNN classifier using the Euclidean distance metric, which achieved an accuracy of 10%.

7 Discussions

This section presents a comprehensive analysis of the use of different distance metrics in image classification. We will compare the experimental results we achieved using different distance metrics, followed by an in-depth error analysis.

7.1 Analysis of the Experimental Results

The experimental results clearly demonstrate that the Wasserstein distance metric, achieved by solving the Monge-Kantorovich problem, yields higher accuracy in classifying handwritten digits when the training and test set sizes are small. The higher accuracy achieved with the Wasserstein distance metric suggests that the Wasserstein distance considers the distribution of pixel intensities across the entire image, which results in a more robust capability in capturing the feature representations when data are sparse in the sample space. Unlike the Euclidean distance, which calculates a direct, pixel-wise distance, the Wasserstein distance provides a more holistic similarity measure. Overall, correlation metric seems to be the most robust measure as it is invariant to location and scale.

7.2 Error Analysis

In addition, we identify the errors made by the KNN classifier and visualized a couple of challenging image pairs and easy image pairs, where the pair of images have low distances or high similarities.

We notice that there are cases where the digits in the image are actually different even though they have a high similarity. For example, in figure 5, the left-hand side shows a digit of "4" and a digit of "9" with a low distance. While the right-hand side shows two digits of "0".

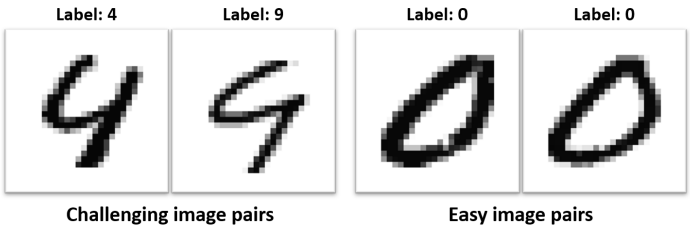

Figure 5: Visualization of challenging and easy image pairs in the MNIST dataset that we found during experiments.

It is interesting to see that although both digits on the left have lower distances, it can be challenging for humans to distinguish them without a closer look.

8 Recommendations

In conclusion, while our project has demonstrated promising results in digit classification in the MNIST dataset using Optimal Transportation theory, several directions remain open for future exploration and improvement. We identify the following potentials for future work: (1)**Colored Images:** Our current focus is on grayscale images. Future research could explore extending the Optimal Transportation framework to color images, introducing additional complexities in probability distribution representation transformations and cost calculations. (2) **Other Cost Functions:** Investigating more complex cost functions beyond the quadratic form could provide deeper insights. Different cost functions could be more suitable for specific types of image data or classification problems. (3) **Other Applications:** Exploring applications beyond digit classification, such as in medical imaging or facial recognition,

helps us to understand the potential of the Nonlinear Monge-Kantorovich Problem in various domains. (4) **Computational Efficiency:** During our experiments, we found that calculating the Wasserstein distance is still time-consuming even after simplification with a numerical solution. This includes exploring more efficient algorithms or parallel processing techniques that could enhance the scalability and speed, making it feasible for larger datasets or real-time applications.

9 Contributions of team members

All authors contributed to the project's problem formulation and initial literature review. They engaged in multiple in-person and online discussions and collaboratively created and rehearsed the presentation. Mingzhen Li and Anthony Hong focused on studying existing literature, conducting mathematical inductions for the 2D Monge-Kantorovich Problem in solutions of primal and dual forms. Jingyuan Zhu and Ruiqi Wang concentrated on applying the solution to the MNIST dataset, analyzing the Gradient Descent Approach, conducting experiments, and performing results and complexity analysis.

References

[1] Ambrosio, Luigi, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*, Springer Science+Business Media, 2005.

[2] An, Dongsheng, Na Lei, Xiaoyin Xu, and Xianfeng Gu. "Efficient optimal transport algorithm by accelerated gradient descent." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 9, pp. 10119-10128. 2022.

[3] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Communications on Pure and Applied Mathematics, Vol. 44, pp. 375-417, 1991.

[4] Evans, Lawrence C. *Partial Differential Equations and Monge–Kantorovich Mass Transfer*, https://math.berkeley.edu/~evans/Monge-Kantorovich.survey.pdf, September, 2001 version.

[5] LeCun, Y., Cortes, C. and Burges, C.J.C. The MNIST Database of Handwritten Digits. New York, USA, 1998.

[6] Michael Snow and Jan Van lent, *Monge's Optimal Transport Distance for Image Classification*, Department of Engineering Design and Mathematics, Centre for Machine Vision, University of the West of England, Bristol, 2018.

[7] Rick Chartrand and Brendt Wohlberg, *A Gradient Descent Solution to the Monge-Kantorovich Problem*, Applied Mathematical Sciences, Vol. 3, no. 22, 1071 - 1080, 2009.

[8] Dvurechensky, P., Gasnikov, A., and Kroshnin, A. Com- putational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In Proceedings of the 35th International Conference on Machine Learning. PMLR, 2018.

[9] Lin, T., Ho, N., and Jordan, M. On efficient optimal trans- port: An analysis of greedy and accelerated mirror de- scent algorithms. In International Conference on Ma- chine Learning, 2019.