NATURAL GRADIENT IN WASSERSTEIN STATISTICAL MANIFOLD

YIFAN CHEN AND WUCHEN LI

Abstract. We study the Wasserstein natural gradient in parametric statistical models with continuous sample space. Our approach is to pull back the L^2-Wasserstein metric tensor in probability density space to parameter space, under which the parameter space become a Riemannian manifold, named the Wasserstein statistical manifold. The gradient flow and natural gradient descent method in parameter space are then derived. When parameterized densities lie in $\mathbb{R}$, we show the induced metric tensor establishes an explicit formula. Computationally, optimization problems can be accelerated by the proposed Wasserstein natural gradient descent, if the objective function is the Wasserstein distance. Examples are presented to demonstrate its effectiveness in several parametric statistical models.

1. Introduction

The statistical distance between probability measures plays an important role in a lot of fields such as data analysis and machine learning, which usually consist in minimizing a loss function as

$$\text{minimize} \quad d(\rho, \rho_e) \quad \text{s.t.} \quad \rho \in \mathcal{P}_\theta.$$

Here $\mathcal{P}_\theta$ is a parameterized subset of the probability density space, and ρ_e is a given target density often referred to an empirical realization of a ground-truth distribution. The function d serves as the distance, which quantifies the difference between densities ρ and ρ_e.

An important example for d is the Kullback-Leibler (KL) divergence, also known as the relative entropy [16], which closely relates to the maximum likelihood estimate in statistics and the field of information geometry [2, 7]. The Hessian operator of KL embeds $\mathcal{P}_\theta$ as a statistical manifold, in which the Riemannian metric is given by the Fisher-Rao metric [34]. Due to Chentsov [14], the Fisher-Rao metric is the only one, up to scaling, that is invariant to statistical embeddings by Markov morphisms. Using Fisher-Rao metric, a natural gradient descent method, realized by a Forward-Euler discretization of the gradient flow in the manifold, has been introduced. It has found many successful applications in a variety of problems such as blind source separation [3] and machine learning [1, 27].

Recently, the Wasserstein distance, introduced through the field of optimal transport, has been attracting increasing attention [32]. One promising property of the Wasserstein distance is its ability to reflect the metric on sample space, rendering it very useful in machine learning [6, 20, 30], statistical models [11, 13] and geophysics [18, 19, 12]. Further, optimal transport theory provides the L^2-Wasserstein metric tensor, which gives the probability density space an infinite-dimensional Riemannian differential structure [21, 24]. The gradient flow with respect to the L^2-Wasserstein metric tensor, known as the Wasserstein gradient flow, have been seen

Key words and phrases. Optimal transport; Information geometry; Wasserstein statistical manifold; Wasserstein Natural gradient.

deep connections to fluid dynamics [10, 31], differential geometry [25] and mean field games [15, 17].

Nevertheless, compared to the Fisher-Rao metric, the Riemannian structure of the Wasserstein metric is mostly investigated in the whole probability space rather than the parameterized subset $\mathcal{P}_\theta$. Therefore, there remains a gap in developing natural gradient concept in a parametric model within the Wasserstein geometry context. Here we are primarily interested in the question whether there exists the Wasserstein metric tensor and the associated Wasserstein natural gradient in a general parameterized subset and whether we can gain computational benefits by considering these structures. We believe the answer to it will serve as a window to bring synergies between the information geometry and optimal transport communities.

In this paper, we embed the Wasserstein geometry to parametric probability models with continuous sample space. Like in [22, 23], we pull back the L^2-Wasserstein metric tensor into parameter space, making it become a finite-dimensional Riemannian manifold. It allows us to derive the constrained Wasserstein gradient flow in parameter space. The discretized version of the flow leads to the Wasserstein natural gradient descent method, in which the induced metric tensor acts as a preconditioning term in the standard gradient descent iteration. When the dimension of densities is one, we obtain an explicit formula of this metric tensor. Precisely, given $\rho(x,\theta)$ as a parameterized density, $x \in \mathbb{R}^1$ and $\theta \in \Theta \subset \mathbb{R}^d$, the L^2-Wasserstein metric tensor on Θ will be

$$G_W(\theta) = \int \frac{1}{\rho(x,\theta)} (\nabla_\theta F(x,\theta))^T \nabla_\theta F(x,\theta) dx,$$

where $F(y,\theta) = \int_{-\infty}^{x} \rho(y,\theta) dy$ is the cumulative distribution function of $\rho(x,\theta)$. We apply the natural gradient descent induced by $G_W(\theta)$ to Wasserstein metric modeled problems. It is seen that the Wasserstein gradient descent outperforms the Euclidean and Fisher-Rao natural gradient descent in the iterations. We give theoretical justifications of this phenomenon by showing that the Wasserstein gradient descent performs asymptotically Newton method in this case. Detailed description of the Hessian matrix is also presented by leveraging techniques in one-dimensional OT.

In literature, there are pioneers in the direction of constrained Wasserstein gradient flow. [10] studies density space with fixed mean and variance. Compared to them, we focus on a density set parameterized by a finite dimensional parameter space. Also, there have been many works linking information geometry and optimal transport [4, 37]. In particular, the Wasserstein metric tensor on Gaussian distributions exhibits explicit form [35], which leads to extensive studies between Wasserstein and Fisher-Rao metric for this model [26, 28, 29, 33]. In contrast to their works, we extend the Wasserstein metric tensor to general parametric models. It allows us to discuss the Wasserstein gradient flow systematically.

This paper is organized as follows. In section 2, we briefly review the theory of optimal transport with a concentration on its Riemannian differential structure. In section 3, we introduce the Wasserstein statistical manifolds by defining the metric tensor in the parameter space directly. The Wasserstein gradient flow and natural gradient descent method are then derived. We give a concise study of the metric tensor for one-dimensional densities, showing its connection to Fisher information matrix. In this case, we theoretically analyze the effect of this natural gradient in Wasserstein metric modeled problems. In section 4, examples are presented to justify the previous discussions.

2. Review of Optimal Transport Theory

In this section, we briefly review the theory of optimal transport (OT). We note that there are several equivalent definitions of OT, ranging from static to dynamic formulations. In this paper, we focus on the dynamic formulation and its induced Riemannian metric tensor in density space.

The optimal transport problem is firstly proposed by Monge in 1781: given two probability densities ρ^0, ρ^1 on $\Omega \subset \mathbb{R}^n$, the goal is to find a transport plan $T : \Omega \to \Omega$ pushing ρ^0 to ρ^1 that minimizes the whole transportation cost, i.e.

$$\inf_T \int_\Omega d\left(x, T(x)\right)\rho^0(x)dx \quad \text{s.t.} \quad \int_A \rho^1(x)dx = \int_{T^{-1}(A)} \rho^0(x)dx, \tag{1}$$

for any Borel subset $A \subset \Omega$. Here function $d\colon \Omega \times \Omega \to \mathbb{R}$ is the ground distance that measures the difference between x and $T(x)$. In the whole discussions we set $d(x, y) = \|x - y\|^2$ as the square of Euclidean distance. We assume all the densities belong to $\mathcal{P}_2(\Omega)$, which is defined as the collection of probability density functions on $\Omega \subset \mathbb{R}^n$ with finite second moment.

In 1942, Kantorovich relaxed the problem to a linear programming:

$$\min_{\pi \in \Pi(\rho^0, \rho^1)} \int_{\Omega \times \Omega} \|x - y\|^2 \pi(x, y) dxdy, \tag{2}$$

where the infimum is taken over the set Π of joint probability measures on $\Omega \times \Omega$ that have marginals ρ^0, ρ^1. This formulation finds a wide array of applications in computation [32].

In recent years, OT connects to a variational problem in density space, known as the Benamou-Brenier formula [8]:

$$\inf_{\Phi_t} \int_0^1 \int_\Omega \|\nabla \Phi(t, x)\|^2 \rho(t, x) dxdt, \tag{3a}$$

where the infimum is taken over the set of Borel potential function $\Phi : [0, 1] \times \Omega \to \mathbb{R}$. Each gradient vector field of potential $\Phi_t = \Phi(t, x)$ on sample space determines a corresponding density path $\rho_t = \rho(t, x)$ as the solution of the continuity equation:

$$\frac{\partial \rho(t, x)}{\partial t} + \nabla \cdot \left(\rho(t, x)\nabla \Phi(t, x)\right) = 0, \quad \rho(0, x) = \rho^0(x), \quad \rho(1, x) = \rho^1(x). \tag{3b}$$

Here $\nabla\cdot$ and ∇ are the divergence and gradient operators in $\mathbb{R}^n$. If Ω is a compact set, the zero flux condition (Neumann condition) is proposed on the boundary of Ω. This is to ensure that $\int_\Omega \frac{\partial \rho(t,x)}{\partial t} dx = 0$, so that the total mass is conserved.

Under mild regularity assumptions, the above three formulations (1) (2) (3) are equivalent, see details in [36]. Their optimal quantity is denoted by $(W_2(\rho^0, \rho^1))^2$, which is called the square of the L^2-Wasserstein distance between ρ^0 and ρ^1. Here the subscript "2" in W_2 indicates that the L^2 ground distance is used. We note that formulation (1) (2) are static, in the sense that only the initial and final states of the transportation are considered. By taking the transportation path into consideration, OT enjoys a dynamical formulation (3). This will be our main interest in the following discussion.

The variational formulation (3) introduces an infinite-dimensional Riemannian structure in density space. For better illustration, suppose Ω is compact and consider the set of smooth and strictly positive densities

$$\mathcal{P}_+(\Omega) = \Big\{\rho \in C^\infty(\Omega) \colon \rho(x) > 0,\ \int_\Omega \rho(x)dx = 1\Big\} \subset \mathcal{P}_2(\Omega).$$

Denote by $\mathcal{F}(\Omega) := C^\infty(\Omega)$ the set of smooth real valued functions on Ω. The tangent space of $\mathcal{P}_+(\Omega)$ is given by

$$T_\rho\mathcal{P}_+(\Omega) = \Big\{\sigma \in \mathcal{F}(\Omega) \colon \int_\Omega \sigma(x)dx = 0\Big\}.$$

Given $\Phi \in \mathcal{F}(\Omega)$ and $\rho \in \mathcal{P}_+(\Omega)$, define

$$V_\Phi(x) := -\nabla\cdot(\rho(x)\nabla\Phi(x)) \in T_\rho\mathcal{P}_+(\Omega).$$

Since ρ is positive in a compact region Ω, the elliptic operator identifies the function Φ on Ω modulo additive constants with the tangent vector V_Φ in $\mathcal{P}_+(\Omega)$. This gives an isomorphism

$$\mathcal{F}(\Omega)/\mathbb{R} \to T_\rho\mathcal{P}_+(\Omega), \quad \Phi \mapsto V_\Phi.$$

Here we treat $T_\rho^*\mathcal{P}_+(\Omega) = \mathcal{F}(\Omega)/\mathbb{R}$ as the smooth cotangent space of $\mathcal{P}_+(\Omega)$. The above facts introduce the L^2-Wasserstein metric tensor on density space:

Definition 1 (L^2-Wasserstein metric tensor)**.** *Define the inner product on the tangent space of positive densities $g_\rho \colon T_\rho\mathcal{P}_+(\Omega) \times T_\rho\mathcal{P}_+(\Omega) \to \mathbb{R}$ by*

$$g_\rho(\sigma_1, \sigma_2) = \int_\Omega \nabla\Phi_1(x)\cdot\nabla\Phi_2(x)\rho(x)dx,$$

where $\sigma_1 = V_{\Phi_1}$, $\sigma_2 = V_{\Phi_2}$ with $\Phi_1(x)$, $\Phi_2(x) \in \mathcal{F}(\Omega)/\mathbb{R}$.

With the inner product specified above, the variational problem (3) becomes a geometric action energy in $(\mathcal{P}_+(\Omega), g_\rho)$. As in Riemannian geometry, the square of distance equals the energy of geodesics, i.e.

$$(W_2(\rho^0,\rho^1))^2 = \inf_{\Phi_t}\ \Big\{\int_0^1 g_{\rho_t}(V_{\Phi_t}, V_{\Phi_t})dt \colon \partial_t\rho_t = V_{\Phi_t},\ \rho(0,x) = \rho^0,\ \rho(1,x) = \rho^1\Big\}.$$

This is exactly the form in (3). In this sense, it explains that the dynamical formulation of OT exhibits the Riemannian structure for density space. In [21], $(\mathcal{P}_+(\Omega), g_\rho)$ is named density manifold. More geometric studies are provided in [24, 22].

We note that the geometric treatment of density space can be extended beyond compact Ω and $\mathcal{P}_+(\Omega)$. Replacing $\mathcal{P}_+(\Omega)$ by $\mathcal{P}_2(\Omega)$ and assuming conditions such as absolute continuity of ρ is satisfied, $(\mathcal{P}_2(\Omega), W_2)$ will become a length space. Interested readers can refer to [5] for analytical results.

3. Wasserstein Natural Gradient

In this section, we study parametric statistical models, which relate to parameterized subsets of the probability space $\mathcal{P}_2(\Omega)$. We pull back the L^2-Wasserstein metric tensor into the parameter space, turning it to be a Riemannian manifold. This consideration allows us to introduce the Riemannian (natural) gradient flow on the parameter spaces, which further leads to a natural gradient descent method in optimization. When densities lie in $\mathbb{R}$, we show that the metric tensor establishes an explicit formula. It acts as a positive and asymptotically-Hessian preconditioner for Wasserstein metric related minimizations.

3.1. **Wasserstein statistical manifold.** We adopt the definition of statistical model from [7]. It is represented by a triple (Ω, Θ, ρ), where $\Omega \subset \mathbb{R}^n$ is the continuous sample space, $\Theta \subset \mathbb{R}^d$ is the statistical parameter space, and ρ is the probability density on Ω parameterized by θ such that $\rho\colon \Theta \to \mathcal{P}_2(\Omega)$ and $\rho = \rho(\cdot, \theta)$. For simplicity we assume Ω is either compact or $\Omega = \mathbb{R}^n$, and each $\rho(\cdot, \theta)$ is positive, smooth with finite second moment. The parameter space is a finite dimensional manifold with metric tensor denoted by $\langle \cdot, \cdot \rangle_\theta$. We also use $\langle \cdot, \cdot \rangle$ to represent the Euclidean inner product in $\mathbb{R}^d$.

The Riemannian metric g_θ on Θ will be the pull-back of $g_{\rho(\cdot,\theta)}$ on $\mathcal{P}_2(\Omega)$. That is, for ξ, $\eta \in T_\theta\Theta$, we have

$$g_\theta(\xi, \eta) := g_{\rho(\cdot,\theta)}(d_\theta\rho(\xi), d_\theta\rho(\eta)),$$

where $d_\theta\rho(\xi) = \langle \nabla_\theta\rho(\cdot, \theta), \xi \rangle_\theta$, $d_\theta\rho(\eta) = \langle \nabla_\theta\rho(\cdot, \theta), \eta \rangle_\theta$. The tensor $g_{\rho(\cdot,\theta)}$ involves the solution of elliptic equations and we make the following assumptions on the statistical model (Ω, Θ, ρ):

Assumption 1. *For the statistical model (Ω, Θ, ρ), one of the following two conditions are satisfied:*

(1) *The sample space Ω is compact, and for each $\xi \in T_\theta(\Theta)$, the elliptic equation*

$$\begin{cases} -\nabla \cdot (\rho(x, \theta)\nabla\Phi(x)) = \langle \nabla_\theta\rho(x, \theta), \xi \rangle_\theta \\ \frac{\partial\Phi}{\partial n}|_{\partial\Omega} = 0 \end{cases}$$

has a smooth solution Φ satisfying

$$\int_\Omega \rho(x, \theta)\|\nabla\Phi(x)\|^2 dx < +\infty. \tag{4}$$

(2) *The sample space $\Omega = \mathbb{R}^n$, and for each $\xi \in T_\theta(\Theta)$, the elliptic equation*

$$-\nabla \cdot (\rho(x, \theta)\nabla\Phi(x)) = \langle \nabla_\theta\rho(x, \theta), \xi \rangle_\theta$$

has a smooth solution Φ satisfying

$$\int_\Omega \rho(x, \theta)\|\nabla\Phi(x)\|^2 dx < +\infty \quad \text{and} \quad \int_\Omega \rho(x, \theta)|\Phi(x)|^2 dx < +\infty. \tag{5}$$

Assumption 1 guarantees the action of "pull-back" we described above is well-defined. The conditions (4) and (5) are used to justify the uniqueness of the solutions and the cancellation of boundary terms when integrating by parts.

Proposition 1. *Under assumption 1, the solution Φ is unique modulo the addition of a spatially-constant function.*

Proof. It suffices to show the equation

$$\nabla \cdot (\rho(x, \theta)\nabla\Phi(x)) = 0 \tag{6}$$

only has the trivial solution $\nabla\Phi = 0$ in the space described in assumption 1.

For case (1), we multiple Φ to (6) and integrate it in Ω. Integration by parts result in

$$\int_\Omega \rho(x, \theta)\|\nabla\Phi(x)\|^2 = 0$$

due to the zero flux condition. Hence $\nabla\Phi = 0$.

For case (2), we denote by $B_R(0)$ the ball in $\mathbb{R}^n$ with center 0 and radius R. Multiply Φ to the equation and integrate in $B_R(0)$:

$$\int_{B_R(0)} \rho(x,\theta)\|\nabla\Phi(x)\|^2 = \int_{\partial B_R(0)} \rho(x,\theta)\Phi(x)(\nabla\Phi(x)\cdot n)dx.$$

By Cauchy-Schwarz inequality we can control the right hand side by

$$|\int_{\partial B_R(0)} \rho(x,\theta)\Phi(x)(\nabla\Phi(x)\cdot n)dx|^2 \le \int_{\partial B_R(0)} \rho(x,\theta)\Phi(x)^2 dx \cdot \int_{\partial B_R(0)} \rho(x,\theta)\|\nabla\Phi(x)\|^2 dx.$$

However, due to (5), there exists a sequence $R_k, k \ge 1$, such that $R_{k+1} > R_k$, $\lim_{k\to+\infty} R_k = \infty$ and

$$\lim_{k\to+\infty} \int_{\partial B_{R_k}(0)} \rho(x,\theta)\Phi(x)^2 dx = \lim_{k\to+\infty} \int_{\partial B_{R_k}(0)} \rho(x,\theta)\|\nabla\Phi(x)\|^2 dx = 0.$$

Hence

$$\lim_{k\to+\infty} |\int_{\partial B_{R_k}(0)} \rho(x,\theta)\Phi(x)(\nabla\Phi(x)\cdot n)dx|^2 = 0,$$

which leads to

$$\int_{\mathbb{R}^n} \rho(x,\theta)\|\nabla\Phi(x)\|^2 = 0.$$

Thus $\nabla\Phi = 0$, which is the trivial solution. □

Since we deal with positive ρ, the existence of solutions to case (1) in assumption 1 is ensured by the theory of elliptic equations. For case (2), i.e. $\Omega = \mathbb{R}^n$, we show when ρ is Gaussian distribution in $\mathbb{R}^d$, the existence of solution Φ is guaranteed and exhibits explicit formulation in our examples. Although we only deal with compact Ω or the whole $\mathbb{R}^n$, the treatment to some other Ω, such as the half space of $\mathbb{R}^n$, is similar and omitted.

Definition 2 (L^2-Wasserstein metric tensor in parameter space)**.** *Under assumption 1, the inner product g_θ on $T_\theta(\Theta)$ is defined as*

$$g_\theta(\xi,\eta) = \int_\Omega \rho(x,\theta)\nabla\Phi_\xi(x)\cdot\nabla\Phi_\eta(x)dx,$$

where ξ,η are tangent vectors in $T_\theta(\Theta)$, Φ_ξ and Φ_η satisfy $\langle\nabla_\theta\rho(x,\theta),\xi\rangle_\theta = -\nabla\cdot(\rho\nabla\Phi_\xi(x))$ and $\langle\nabla_\theta\rho(x,\theta),\eta\rangle_\theta = -\nabla\cdot(\rho\nabla\Phi_\eta(x))$.

Generally, (Θ, g_θ) will be a Pseudo-Riemannian manifold. However, if the statistical model is non-degenerate, i.e., g_θ is positive definite on the tangent space $T_\theta(\Theta)$, then (Θ, g_θ) forms a Riemannian manifold. We call (Θ, g_θ) the Wasserstein statistical manifold.

Proposition 2. *The metric tensor can be written as*

$$g_\theta(\xi,\eta) = \xi^T G_W(\theta)\eta, \tag{7}$$

where $G_W(\theta) \in \mathbb{R}^{d\times d}$ is a positive definite matrix and can be represented by

$$G_W(\theta) = G_\theta^T A(\theta) G_\theta,$$

in which $A_{ij}(\theta) = \int_\Omega \partial_{\theta_i}\rho(x,\theta)(-\Delta_\theta)^{-1}\partial_{\theta_j}\rho(x,\theta)dx$ and $-\Delta_\theta = -\nabla\cdot(\rho(x,\theta)\nabla)$. The matrix G_θ associates with the original metric tensor in Θ such that $\langle\dot\theta_1,\dot\theta_2\rangle_\theta = \dot\theta_1^T G_\theta \dot\theta_2$ for any $\dot\theta_1,\dot\theta_2 \in T_\theta(\Theta)$. If Θ is Euclidean space then $G_W(\theta) = A(\theta)$.

Proof. Write down the metric tensor

$$\begin{aligned} g_\theta(\xi,\eta) &= \int_\Omega \rho(x,\theta)\nabla\Phi_\xi(x)\cdot\nabla\Phi_\eta(x)dx \\ &\overset{a)}{=} \int_\Omega \langle\nabla_\theta\rho(x,\theta),\xi\rangle_\theta\cdot\Phi_\eta(x)dx \\ &= \int_\Omega \langle\nabla_\theta\rho(x,\theta),\xi\rangle_\theta(-\Delta_\theta)^{-1}\langle\nabla_\theta\rho(x,\theta),\eta\rangle_\theta dx \end{aligned}$$

where $a)$ is due to integration by parts. Comparing the above equation with (7) finishes the proof. □

Given this $G_W(\theta)$, we derive the geodesic in this manifold and illustrate its connection to the geodesic in $\mathcal{P}_2(\Omega)$ as follows.

Proposition 3. *The geodesics in* (Θ, g_θ) *satisfies*

$$\begin{cases} \dot\theta - G_W(\theta)^{-1}S = 0 \\ \dot S + \frac{1}{2}S^T\frac{\partial}{\partial\theta}G_W(\theta)^{-1}S = 0 \end{cases} \tag{8}$$

Proof. In geometry, the square of geodesic distance d_W between $\rho(\cdot,\theta^0)$ and $\rho(\cdot,\theta^1)$ equals the energy functional:

$$d_W^2(\rho^0(\cdot,\theta),\rho^1(\cdot,\theta)) = \inf_{\theta(t)\in C^1(0,1)}\{\int_0^1 \dot\theta(t)^T G_W(\theta)\dot\theta(t)dt\colon \theta(0)=\theta^0, \theta(1)=\theta^1\}. \tag{9}$$

The minimizer of (9) satisfies the geodesic equation. Let us write down the Lagrangian $L(\dot\theta,\theta) = \frac{1}{2}\dot\theta^T G_W(\theta)\dot\theta$. The geodesic satisfies the Euler-Lagrange equation

$$\frac{d}{dt}\nabla_{\dot\theta}L(\dot\theta,\theta) = \nabla_\theta L(\dot\theta,\theta).$$

By the Legendre transformation,

$$H(S,\theta) = \sup_{\dot\theta\in T_\theta(\Theta)} S^T\dot\theta - L(\dot\theta,\theta).$$

Then $S = G_W(\theta)\dot\theta$ and $H(S,\theta) = \frac{1}{2}S^T G_W(\theta)^{-1}S$. Thus we derive the Hamilton's equations

$$\dot\theta = \partial_S H(\theta,S), \quad \dot S = -\partial_\theta H(\theta,S).$$

This recovers (8). □

Remark 1. *We recall the Wasserstein geodesic equation in* $(\mathcal{P}_2(\Omega), W_2)$*:*

$$\begin{cases} \frac{\partial\rho(t,x)}{\partial t} + \nabla\cdot(\rho(t,x)\nabla\Phi(t,x)) = 0 \\ \frac{\partial\Phi(t,x)}{\partial t} + \frac{1}{2}(\nabla\Phi(t,x))^2 = 0 \end{cases}$$

The above PDE pair contains both continuity equation and Hamilton-Jacobi equation. Our equation (8) can be viewed as the continuity equation and Hamilton-Jacobi equation on the parameter space. The difference is that when restricted to a statistical model, the continuity equation and the associated Hamilton-Jacobi equation can only flow in the probability densities constrained in $\rho(\Theta)$.

Remark 2. *If the optimal flow $\rho_t, 0 \leq t \leq 1$ in the continuity equation* (3b) *totally lies in the probability subspace parameterized by θ, then the two geodesic distances coincide:*

$$d_W(\theta^0, \theta^1) = W_2(\rho^0(\cdot, \theta), \rho^1(\cdot, \theta)).$$

It is well known that the optimal transportation path between two Gaussian distributions will also be Gaussian distributions. Hence when $\rho(\cdot, \theta)$ are Gaussian measures, the above condition is satisfied. This means Gaussian is a totally geodesic submanifold. In general, d_W will be different from the L^2 Wasserstein metric. We will demonstrate this fact in our numerical examples.

3.2. **Wasserstein natural gradient.** Based on the Riemannian structure established in the previous section, we are able to introduce the gradient flow on the parameter space (Θ, g_θ). Given an objective function $R(\theta)$, the associated gradient flow will be:

$$\frac{d\theta}{dt} = -\nabla_g R(\theta).$$

Here ∇_g is the Riemannian gradient operator satisfying

$$g_\theta(\nabla_g R(\theta), \xi) = \nabla_\theta R(\theta) \cdot \xi$$

for any tangent vector $\xi \in T_\theta \Theta$, where ∇_θ represents the Euclidean gradient operator.

Proposition 4. *The gradient flow of function $R \in C^1(\Theta)$ in (Θ, g_θ) satisfies*

$$\frac{d\theta}{dt} = -G_W(\theta)^{-1} \nabla_\theta R(\theta). \tag{10}$$

Proof. By the definition of gradient operator,

$$\nabla_g R(\theta)^T G_W(\theta) \xi = \nabla_\theta R(\theta) \cdot \xi,$$

for any ξ. Thus $\nabla_g R(\theta) = G_W(\theta)^{-1} \nabla_\theta R(\theta)$. □

When $R(\theta) = R(\rho(\cdot, \theta))$, i.e. the function is implicitly determined by the density $\rho(\cdot, \theta)$, the Riemannian gradient can naturally reflect the change in the probability density domain. This will be expressed in our experiments by using Forward-Euler to solve the gradient flow numerically. The iteration writes

$$\theta^{n+1} = \theta^n - \tau G_W(\theta^n)^{-1} \nabla_\theta R(\rho(\cdot, \theta^n)). \tag{11}$$

This iteration of θ^{n+1} can also be understood as an approximate solution to the following problem:

$$\arg\min_\theta \; R(\rho(\cdot, \theta)) + \frac{d_W(\rho(\cdot, \theta^n), \rho(\cdot, \theta))^2}{2\tau}.$$

The approximation goes as follows. Note the Wasserstein metric tensor satisfies

$$d_W(\rho(\cdot, \theta + \Delta\theta), \rho(\cdot, \theta))^2 = \frac{1}{2}(\Delta\theta)^T G_W(\theta)(\Delta\theta) + o\left((\Delta\theta)^2\right) \quad \text{as} \quad \Delta\theta \to 0,$$

and $R(\rho(\cdot, \theta + \Delta\theta)) = R(\rho(\cdot, \theta)) + \langle \nabla_\theta R(\rho(\cdot, \theta)), \Delta\theta \rangle + O((\Delta\theta)^2)$. Ignoring high-order items we obtain

$$\theta^{n+1} = \arg\min_\theta \; \langle \nabla_\theta R(\rho(\cdot, \theta^n)), \theta - \theta^n \rangle + \frac{(\theta - \theta^n)^T G_W(\theta^n)(\theta - \theta^n)}{2\tau}.$$

This recovers (11). It explains (11) is the steepest descent with respect to the change of probability distributions measured by W_2.

In fact, (11) shares the same spirit in natural gradient [1] with respect to Fisher-Rao metric. To avoid ambiguity, we call it the Fisher-Rao natural gradient. It considers θ^{n+1} as an approximate solution of

$$\arg\min_{\theta} \; R(\rho(\cdot,\theta)) + \frac{D_{\mathrm{KL}}(\rho(\cdot,\theta)\|\rho(\cdot,\theta^n))}{\tau},$$

where D_{KL} represents the Kullback-Leibler divergence, i.e. given two densities p, q on Ω, then

$$D_{\mathrm{KL}}(p\|q) = \int_\Omega p(x)\log(\frac{p(x)}{q(x)})dx.$$

In our case, we replace the KL divergence by the constrained Wasserstein metric. For this reason, we call (11) the Wasserstein natural gradient descent method.

3.3. **1D densities.** In the following we concentrate on the one dimensional sample space, i.e. $\Omega = \mathbb{R}$. We show that g_θ exhibits an explicit formula. From it, we demonstrate that when the minimization is modeled by Wasserstein distance, namely $R(\rho(\cdot,\theta))$ is related to W_2, then $G_W(\theta)$ will approach the Hessian matrix of $R(\rho(\cdot,\theta))$ at the minimizer.

Proposition 5. *Suppose $\Omega = \mathbb{R}, \Theta = \mathbb{R}^d$ is the Euclidean space, and assumption 1 is satisfied, then the Riemannian inner product on the Wasserstein statistical manifold (Θ, g_θ) has explicit form*

$$G_W(\theta) = \int_{\mathbb{R}} \frac{1}{\rho(x,\theta)}(\nabla_\theta F(x,\theta))^T \nabla_\theta F(x,\theta)dx, \tag{12}$$

such that $g_\theta(\xi,\eta) = \langle \xi, G_W(\theta)\eta\rangle$.

Proof. When $\Omega = \mathbb{R}$, we have

$$g_\theta(\xi,\eta) = \int_{\mathbb{R}} \rho(x,\theta)\Phi'_\xi(x)\cdot\Phi'_\eta(x)dx,$$

where $\langle\nabla_\theta\rho(x,\theta),\xi\rangle = \Big(\rho\Phi'_\xi(x)\Big)'$ and $\langle\nabla_\theta\rho(x,\theta),\eta\rangle = \big(\rho\Phi'_\eta(x)\big)'$.
Integrating the two sides yields

$$\int_{-\infty}^{y} \langle\nabla_\theta\rho(x,\theta),\xi\rangle = \rho\Phi'_\xi(y).$$

Denote by $F(y) = \int_{-\infty}^{y} \rho(x)dx$ the cumulative distribution function of ρ, then

$$\Phi'_\xi(x) = \frac{1}{\rho(x,\theta)}\langle\nabla_\theta F(x,\theta),\xi\rangle,$$

and

$$g_\theta(\xi,\eta) = \int_{\mathbb{R}} \frac{1}{\rho(x,\theta)}\langle\nabla_\theta F(x,\theta),\xi\rangle\langle\nabla_\theta F(x,\theta),\eta\rangle dx.$$

This means $g_\theta(\xi,\eta) = \langle\xi, G_W(\theta)\eta\rangle$ and we obtain (12). Since assumption 1 is satisfied, this integral is well-defined. $\square$

Recall the Fisher-Rao metric tensor, also known as the Fisher information matrix:

$$\begin{aligned} G_F(\theta) &= \int_{\mathbb{R}} \rho(x,\theta)(\nabla_\theta \log \rho(x,\theta))^T \nabla_\theta \log \rho(x,\theta) dx \\ &= \int_{\mathbb{R}} \frac{1}{\rho(x,\theta)} (\nabla_\theta \rho(x,\theta))^T \nabla_\theta \rho(x,\theta) dx, \end{aligned}$$

where we use the fact $\nabla_\theta \log \rho(x,\theta) = \frac{1}{\rho(x,\theta)} \nabla_\theta \rho(x,\theta)$. Compared to the Fisher-Rao metric tensor, our Wasserstein metric tensor $G_W(\theta)$ only changes the density function in the integral to the corresponding cumulative distribution function. We note the condition that ρ is everywhere positive can be relaxed, for example, by assuming each component of $\nabla_\theta F(x,\theta)$, when viewed as a density in $\mathbb{R}$, is absolutely continuous with respect to $\rho(x,\theta)$. Then we can use the associated Radon-Nikodym derivative to define the integral. This treatment is similar to the one for Fisher-Rao metric tensor [7].

Now we turn to study the computational property of natural gradient method with Wasserstein metric. For standard Fisher-Rao natural gradient, it is known that when $R(\rho(\cdot,\theta)) = \mathrm{KL}(\rho(\cdot,\theta), \rho(\cdot,\theta^*))$, then

$$\lim_{\theta \to \theta^*} G_F(\theta) = \nabla^2_\theta R(\rho(\cdot,\theta^*)).$$

Hence, $G_F(\theta)$ will approach the Hessian of R at the minimizer. Regarding this, the Fisher-Rao natural gradient descent iteration

$$\theta^{n+1} = \theta^n - G_F(\theta^n)^{-1} \nabla_\theta R(\rho(\cdot,\theta)),$$

will be asymptotically Newton method for KL divergence related minimization.

We would like to demonstrate the similar result for Wasserstein natural gradient. In other words, we shall show the Wasserstein natural gradient will be asymptotically Newton method for Wasserstein distance related minimization. To achieve this, we start by proving a detailed description of the Hessian matrix for the Wasserstein metric in Theorem 1. Throughout the following discussion, we use the notation $T'(x,\theta)$ to represent the derivative of T with respect to the x variable. We first make the following assumption which is needed in the proof of Theorem 1 to interchange the differentiation and integration.

Assumption 2. *For any $\theta_0 \in \Theta$, there exists a neighborhood $N(\theta_0) \subset \Theta$, such that*

$$\begin{aligned} &\int_\Omega \max_{\theta \in N(\theta_0)} \left|\frac{\partial^2 F(x,\theta)}{\partial\theta_i \partial\theta_j}\right| \, dx < +\infty \\ &\int_\Omega \max_{\theta \in N(\theta_0)} \left|\frac{\partial \rho(x,\theta)}{\partial\theta_i}\right| \, dx < +\infty \\ &\int_\Omega \max_{\theta \in N(\theta_0)} \frac{1}{\rho(x,\theta)} \left|\frac{\partial F(x,\theta)}{\partial\theta_i} \frac{\partial F(x,\theta)}{\partial\theta_j}\right| dx < +\infty \end{aligned}$$

for each $1 \le i, j \le d$.

Theorem 1. *Consider the statistical model (Ω, Θ, ρ), in which $\rho(\cdot,\theta)$ is positive and Ω is a compact region in $\mathbb{R}$. Suppose assumption 1 and 2 are satisfied and $T'(x,\theta)$ is uniformly bounded for all x when θ is fixed. If the objective function has the form*

$$R(\rho(\cdot,\theta)) = \frac{1}{2} \left(W_2(\rho(\cdot,\theta), \rho^*)\right)^2,$$

where ρ^* is the ground truth density, then

$$\nabla_\theta^2 R(\rho(\cdot,\theta)) = \int_\Omega (T(x,\theta)-x)\nabla_\theta^2 F(x,\theta)dx + \int_\Omega \frac{T'(x,\theta)}{\rho(x,\theta)}(\nabla_\theta F(x,\theta))^T \nabla_\theta F(x,\theta)dx, \quad (13)$$

in which $T(\cdot,\theta)$ is the optimal transport map between $\rho(\cdot,\theta)$ and ρ^*, the function $F(\cdot,\theta)$ is the cumulative distribution function of $\rho(\cdot,\theta)$.

Proof. We recall the three formulations of OT in section 2 and the following facts for 1D Wasserstein distance. They will be used in the proof.

(i) When $\Omega \subset \mathbb{R}$, the optimal map will have explicit formula, namely $T(x) = F_1^{-1}(F_0(x))$, where F_0, F_1 are cumulative distribution functions of ρ^0, ρ^1 respectively. Moreover, T satisfies

$$\rho^0(x) = \rho^1(T(x))T'(x). \quad (14)$$

(ii) The dual of linear programming (2) has the form

$$\max_\phi \int_\Omega \phi(x)\rho^0(x)dx + \int_\Omega \phi^c(x)\rho^1(x)dx, \quad (15)$$

in which ϕ and ϕ^c satisfy

$$\phi^c(y) = \inf_{x\in\Omega} \|x-y\|^2 - \phi(x).$$

(iii) We have the relation $\nabla\phi(x) = 2(x - T(x))$ for the optimal T and ϕ.

Using the above three facts, we have

$$R(\rho(\cdot,\theta)) = \frac{1}{2}\left(W_2(\rho(\cdot,\theta),\rho^*)\right)^2 = \frac{1}{2}\int_\Omega |x - F_*^{-1}(F(x,\theta))|^2 \rho(x,\theta)dx,$$

where F_* is the cumulative distribution function of ρ^*. We first compute $\nabla_\theta R(\rho(\cdot,\theta))$. Fix θ and assume the dual maximum is achieved by ϕ^* and ϕ^{c*}:

$$R(\rho(\cdot,\theta)) = \frac{1}{2}\int_\Omega \phi(x)^*\rho(x,\theta)dx + \int_\Omega \phi^{c*}(x)\rho^*(x)dx.$$

Then, for any $\hat\theta \in \Theta$,

$$R(\rho(\cdot,\hat\theta)) \le \frac{1}{2}\int_\Omega \phi^*(x)\rho(x,\hat\theta)dx + \int_\Omega \phi^{c*}(x)\rho^*(x)dx,$$

and the equality holds when $\hat\theta = \theta$. Thus

$$\nabla_\theta R(\rho(\cdot,\theta)) = \nabla_\theta \frac{1}{2}\int_\Omega \phi^*(x)\rho(x,\theta)dx = \frac{1}{2}\int_\Omega \phi(x,\theta)\nabla_\theta\rho(x,\theta)dx,$$

in which integration and differentiation are interchangeable due to Lebesgue dominated convergence theorem and assumption 2. The function $\phi(x,\theta)$ is the Kantorovich potential associated with $\rho(x,\theta)$.

As is mentioned in (iii), $(\phi(x,\theta))' = 2(x - T(x,\theta))$, which leads to

$$\nabla_\theta R(\rho(\cdot,\theta)) = -\int_\Omega (x - T(x,\theta))\nabla_\theta F(x,\theta)dx.$$

Differentiation with respect to θ and interchange integration and differentiation:

$$\nabla_\theta^2 R(\rho(\cdot,\theta)) = -\int_\Omega (x - T(x,\theta))\nabla_\theta^2 F(x,\theta)dx + \int_\Omega (\nabla_\theta T(x,\theta))^T \nabla_\theta F(x,\theta)dx.$$

On the other hand, T satisfies the following equation as in (14):

$$\rho(x,\theta) = \rho^*(T(x,\theta))T'(x,\theta). \tag{16}$$

Differentiating with respect to θ and noticing that the derivative of the right hand side has a compact form:

$$\nabla_\theta \rho(x,\theta) = \left(\rho^*(T(x,\theta))\nabla_\theta T(x,\theta)\right)',$$

and hence

$$\nabla_\theta T(x,\theta) = \frac{1}{\rho^*(T(x,\theta))}\int_{-\infty}^{x} \nabla_\theta \rho(y,\theta)dy = \frac{\nabla_\theta F(x,\theta)}{\rho^*(T(x,\theta))}.$$

Combining them together we obtain

$$\nabla_\theta^2 R(\rho(\cdot,\theta)) = \int_\Omega (T(x,\theta)-x)\nabla_\theta^2 F(x,\theta)dx + \int_\Omega \frac{1}{\rho^*(T(x,\theta))}(\nabla_\theta F(x,\theta))^T \nabla_\theta F(x,\theta)dx. \tag{17}$$

Substituting $\rho^*(T(x,\theta))$ by $\rho(x,\theta)$ and $T(x,\theta)$ based on (16), we obtain (13). Since assumption 1 and 2 are satisfied, and $T'(x,\theta)$ is uniformly bounded, the integral in (13) is well-defined. □

Proposition 6. *Under the condition in Theorem 1, if the ground-truth density satisfies $\rho^* = \rho(\cdot,\theta^*)$ for some $\theta^* \in \Theta$, then*

$$\lim_{\theta\to\theta^*} \nabla_\theta^2 R(\rho(\cdot,\theta)) = \int_{\mathbb{R}} \frac{1}{\rho(x,\theta^*)}(\nabla_\theta F(x,\theta^*))^T \nabla_\theta F(x,\theta^*)dx = G_W(\theta^*).$$

Proof. When θ approaches θ^*, $T(x,\theta)-x$ will go to zero and T' will go to the identity. We finish the proof by the result in Theorem 1. □

Proposition 6 explains that the Wasserstein natural gradient descent is asymptotically Newton method for Wasserstein related minimization. The preconditioner $G_W(\theta)$ equals to the Hessian matrix of R at ground truth. Moreover, the formula (13) contains more information than proposition 6, and they can be used to find suitable Hessian-like preconditioners. For example, we can see the second term in (13) is different from $G_W(\theta)$ when $\theta \neq \theta^*$. It seems more accurate to use the term

$$\bar{G}_W(\theta) := \int \frac{T'(x,\theta)}{\rho(x,\theta)}(\nabla_\theta F(x,\theta))^T \nabla_\theta F(x,\theta)dx \tag{18}$$

to approximate the Hessian of $R(\rho(\cdot,\theta))$. When θ is near to θ^*, $\bar{G}_W$ is likely to achieve slightly faster convergence to $\nabla_\theta^2 R(\rho(\cdot,\theta^*))$ than G_W. However, the use of $\bar{G}_W(\theta)$ could also have several difficulties. The presence of $T'(x,\theta)$ limits its application to general minimization problem in space Θ which does not involve a Wasserstein metric objective function. Also, the computation of $T'(x,\theta)$ might suffer from potential numerical instability, especially when T is not smooth, which is often the case when the ground-truth density is a sum of delta functions. This fact is also expressed in our numerical examples.

4. Examples

In this section, we consider several concrete statistical models. We compute the related metric tensor $G_W(\theta)$, either explicitly or numerically, and further calculate the geodesic in the Wasserstein statistical manifold. Moreover, we test the Wasserstein natural gradient descent method in the Wasserstein distance based inference and fitting problems [9]. We show that the preconditioner $G_W(\theta)$ exhibits promising performance, leading to stable and fast convergence of the iteration. We also compare our results with the Fisher-Rao natural gradient.

4.1. **Gaussian measures.** We consider the multivariate Gaussian densities $\mathcal{N}(\mu, \Sigma)$ in $\mathbb{R}^n$:

$$\rho(x, \theta) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right),$$

where $\theta = (\mu, \Sigma) \in \Theta := \mathbb{R}^n \times \mathrm{Sym}^+(n, \mathbb{R})$. Here $\mathrm{Sym}^+(n, \mathbb{R})$ is the $n \times n$ positive symmetric matrix group. We also denote $\mathrm{Sym}(n, \mathbb{R})$ the $n \times n$ symmetric matrix group. We obtain an explicit formula for g_θ by using definition 2.

Proposition 7. *The Wasserstein metric tensor for the multivariate Gaussian model is*

$$g_\theta\left(\xi, \eta\right) = \langle \dot{\mu_1}, \dot{\mu_2} \rangle + tr(S_1\Sigma S_2),$$

for any $\xi, \eta \in T_\theta\Theta$. Here $\xi = (\dot{\mu}_1, \dot{\Sigma}_1)$ and $\eta = (\dot{\mu}_2, \dot{\Sigma}_2)$, in which $\dot{\mu}_1, \dot{\mu}_2 \in \mathbb{R}^n, \dot{\Sigma}_1, \dot{\Sigma}_2 \in$ Sym$(n, \mathbb{R})$, and the symmetric matrix S_1, S_2 satisfy $\dot{\Sigma}_1 = S_1\Sigma + \Sigma S_1, \dot{\Sigma}_2 = S_2\Sigma + \Sigma S_2$.

Proof. First we examine the elliptic equation in definition 2 has the solution explained in Proposition 7. Write down the equation

$$\langle \nabla_\theta \rho(x, \theta), \xi \rangle_\theta = -\nabla \cdot (\rho(x, \theta) \nabla \Phi_\xi(x)).$$

By some computations we have

$$\begin{aligned}
&\langle \nabla_\theta \rho(x, \theta), \xi \rangle_\theta = \langle \nabla_\mu \rho(x, \theta), \dot{\mu}_1 \rangle + \mathrm{tr}(\nabla_\Sigma \rho(x, \theta) \dot{\Sigma}_1), \\
&\langle \nabla_\mu \rho(x, \theta), \dot{\mu}_1 \rangle = \dot{\mu}_1^T \Sigma^{-1}(x-\mu) \cdot \rho(x, \theta), \\
&\mathrm{tr}(\nabla_\Sigma \rho(x, \theta)\dot{\Sigma}_1) = -\frac{1}{2}\left(\mathrm{tr}(\Sigma^{-1}\dot{\Sigma}_1) - (x-\mu)^T\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}(x-\mu)\right) \cdot \rho(x, \theta), \\
&-\nabla \cdot (\rho \nabla \Phi_\xi(x)) = (\nabla \Phi_\xi(x))\Sigma^{-1}(x-\mu) \cdot \rho(x, \theta) - \rho(x, \theta)\Delta\Phi_\xi(x).
\end{aligned}$$

Observing these equations, we let $\nabla\Phi_\xi(x) = (S_1(x-\mu) + \dot{\mu}_1)^T$, and $\Delta\Phi_\xi = \mathrm{tr}(S_1)$, where S_1 is a symmetric matrix to be determined. By comparison of the coefficients and the fact $\dot{\Sigma}_1$ is symmetric, we obtain

$$\dot{\Sigma}_1 = S_1\Sigma + \Sigma S_1.$$

Similarly $\nabla\Phi_\eta(x) = (S_2(x-\mu) + \dot{\mu}_2)^T$ and

$$\dot{\Sigma}_2 = S_2\Sigma + \Sigma S_2.$$

Then

$$\begin{aligned}
g_\theta(\xi, \eta) &= \int \rho(x, \theta) \nabla\Phi_\xi(x) \cdot \nabla\Phi_\eta(x) dx \\
&= \langle \dot{\mu_1}, \dot{\mu_2} \rangle + \mathrm{tr}(S_1\Sigma S_2).
\end{aligned}$$

It is easy to check Φ satisfies the condition (5) and the uniqueness is guaranteed. □

For Gaussian distributions, the above derived metric tensor has already been revealed in [35, 26]. Our calculation shows that it is a particular formulation of G_W. In the following, we turn to several one-dimensional non-Gaussian distributions and illustrate the metric tensor and the related geodesics, gradient flow numerically.

4.2. **Mixture model.** We consider a generalized version of the Gaussian distribution, namely the Gaussian mixture model. For simplicity we assume there are two components, i.e. $a\mathcal{N}(\mu_1,\sigma_1)+(1-a)\mathcal{N}(\mu_2,\sigma_2)$ with density functions:

$$\rho(x,\theta)=\frac{a}{\sigma_1\sqrt{2\pi}}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}+\frac{1-a}{\sigma_2\sqrt{2\pi}}e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}},$$

where $\theta=(a,\mu_1,\sigma_1^2,\mu_2,\sigma_2^2)$ and $a\in[0,1]$.

We first compute the geodesic of Wasserstein statistical manifold numerically. Set $\theta^0=(0.3,-3,0.5^2,-5,0.4^2)$ and $\theta^1=(0.6,7,0.4^2,5,0.3^2)$. Their density functions are shown in Figure 1:

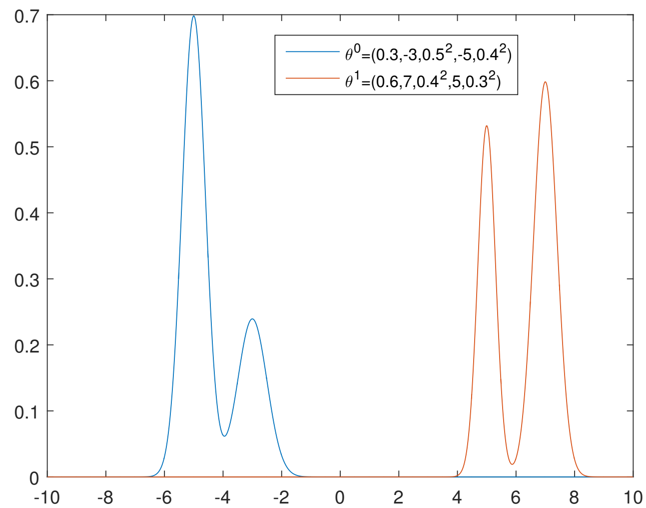

FIGURE 1. Densities of Gaussian mixture distribution

To compute the geodesic in the Wasserstein statistical manifold, we solve the optimal control problem in (9) numerically via a direct method. Discretize the problem as

$$\min_{\theta_i,1\le i\le N-1} N\sum_{i=0}^{N-1}(\theta_{i+1}-\theta_i)^T G_W(\theta_i)(\theta_{i+1}-\theta_i),$$

where $\theta_0=\theta^0,\theta_N=\theta^1$ and the discrete time step-size is $1/N$. We use coordinate descent method, i.e. applying gradient on each $\theta_i, 1\le i\le N-1$ alternatively till convergence.

The geodesic in the whole density space is obtained by first computing the optimal transportation map T, using the explicit formula in one dimension $Tx=F_1^{-1}(F_0(x))$. Here F_0,F_1 are the cumulative distribution functions of ρ^0 and ρ^1. Then the geodesic probability densities satisfies $\rho(t,x)=(tT+(1-t)I)\#\rho_0(x)$ for $0\le t\le 1$, where $\#$ is the push forward operator. The result is shown in Figure 2.

Figure 2 demonstrates that the geodesic in the whole density manifold does not lie in the sub-manifold formed by the mixture distribution, and thus the distance d_W differs from the L^2 Wasserstein metric. Hence the optimal transport in the whole density space destroys the

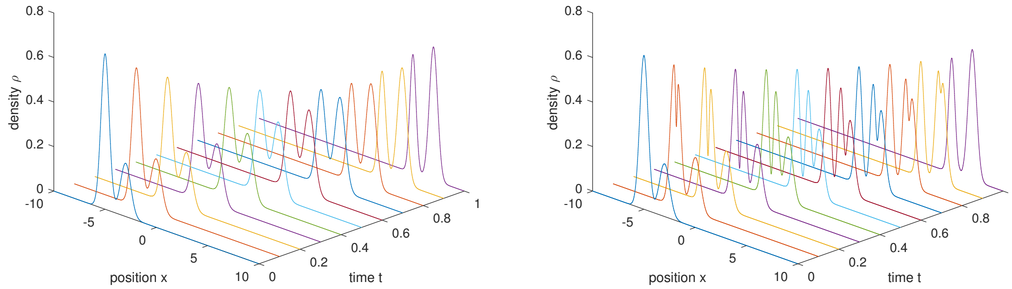

FIGURE 2. Geodesic of Gaussian mixtures; left: in the Wasserstein statistical manifold; right: in the whole density space

geometric shape during its path, which is not a desired property when we perform transportation.

Next, we test the Wasserstein natural gradient method in optimization. Consider the Gaussian mixture fitting problem: given N data points $\{x_i\}_{i=1}^N$ obeying the distribution $\rho(x;\theta^1)$ (unknown), we want to infer θ^1 by using these data points, which leads to a minimization as:

$$\min_\theta d\left(\rho(\cdot;\theta), \frac{1}{N}\sum_{i=1}^N \delta_{x_i}(\cdot)\right),$$

where d are certain distance functions on probability space. If we set d to be KL divergence, then the problem will correspond to the maximum likelihood estimate. Since $\rho(x;\theta^1)$ has small compact support, using KL divergence is risky and needs very good initial guess and careful optimization. Here we use the L^2 Wasserstein metric instead and set $N=10^3$. We truncate the distribution in $[-r,r]$ for numerical computation. We choose $r=15$. The Wasserstein metric is effectively computed by the explicit formula

$$\frac{1}{2}\left(W_2(\rho(\cdot;\theta), \frac{1}{n}\sum_{i=1}^N \delta_{x_i})\right)^2 = \frac{1}{2}\int_{-r}^r |x-T(x)|^2\rho(x;\theta)dx,$$

where $T(x) = F_{em}^{-1}(F(x))$ and F_{em} is the cumulative distribution function of the empirical distribution. The gradient with respect to θ can be computed through

$$\nabla_\theta(\frac{1}{2}W^2) = \int_{-r}^r \phi(x)\nabla_\theta\rho(x;\theta)dx,$$

where $\phi(x) = \int_{-r}^x (y-T(y))dy$ is the Kantorovich potential. The derivative $\nabla_\theta\rho(x;\theta)$ is obtained by numerical differentiation. We perform the following five iterative algorithms to solve the optimization problem:

$$\text{Gradient descent (GD)}: \quad \theta_{n+1} = \theta_n - \tau \nabla_\theta (\frac{1}{2}W^2)|_{\theta_n}$$

$$\text{GD with diag-preconditioning}: \quad \theta_{n+1} = \theta_n - \tau P^{-1} \nabla_\theta (\frac{1}{2}W^2)|_{\theta_n}$$

$$\text{Wasserstein GD}: \quad \theta_{n+1} = \theta_n - \tau G_W(\theta_n)^{-1} \nabla_\theta (\frac{1}{2}W^2)|_{\theta_n}$$

$$\text{Modified Wasserstein GD}: \quad \theta_{n+1} = \theta_n - \tau \left(\bar{G}_W(\theta_n)\right)^{-1} \nabla_\theta (\frac{1}{2}W^2)|_{\theta_n}$$

$$\text{Fisher-Rao GD}: \quad \theta_{n+1} = \theta_n - \tau G_F(\theta_n)^{-1} \nabla_\theta (\frac{1}{2}W^2)|_{\theta_n}$$

We consider the diagonal preconditioning because the scale of parameter a is very different from μ_i, σ_i, $1 \leq i \leq 2$. The diagonal matrix P is set to be $\mathrm{diag}(40, 1, 1, 1, 1)$. We choose the initial step-size $\tau = 1$ with line search such that the objective value is always decreasing. The initial guess $\theta = \theta^0$. Below Figure 3 shows the experimental results.

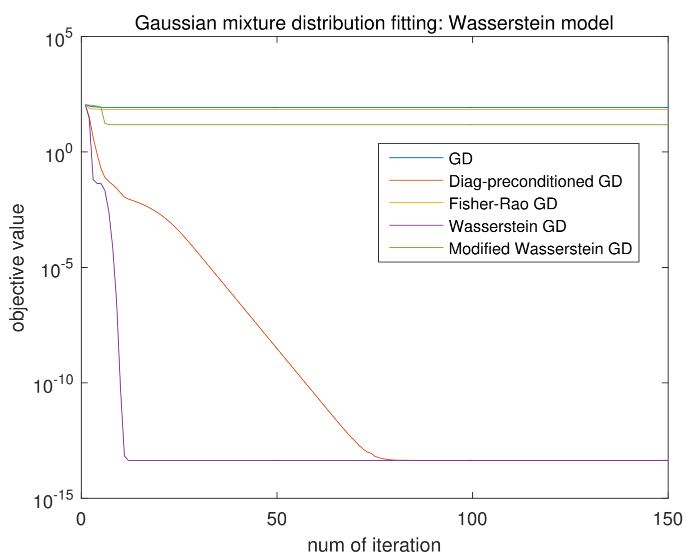

FIGURE 3. objective value

From the figure, it is seen that the Euclidean gradient descent fails to converge. We observed that during iterations the parameter a goes very fast to 1 and then stop updating anymore. This is due to the ill-conditioned nature of the problem, in the sense that the scale of parameter differs drastically. If we use the diagonal matrix P to perform preconditioning, then it converges after approximately 70 steps. If we use Wasserstein gradient descent, then the iterations converge very efficiently, taking less than 10 steps. This demonstrates that $G_W(\theta)$ is well suited for the Wasserstein metric minimization problems, exhibiting very stable behavior. It can automatically detect the parameter scale and the underlying geometry. As a comparison, Fisher-Rao gradient descent fails, which implies $G_F(\theta)$ is not suitable for this Wasserstein metric modeled minimization.

The modified Wasserstein gradient descent does not converge because of the numerical instability of T' and further $\bar{G}_W(\theta)$ in the computation. In the next example with lower dimensional parameter space, however, we will see that $\bar{G}_W(\theta)$ performs better than $G_W(\theta)$. This implies $\bar{G}_W(\theta)$, if computed accurately, might achieve smaller approximation error to the Hessian matrix. Nevertheless, the difference is very slight, and since the matrix $\bar{G}_W(\theta)$ can only be applied to the Wasserstein modeled problem, we tend to believe that $G_W(\theta)$ is a better preconditioner.

4.3. **Gamma distribution.** Consider gamma distribution $\Gamma(\alpha, \beta)$, which has the probability density function

$$\rho(x; \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}.$$

Set $\theta = (\alpha, \beta)$ and $\theta^0 = (2, 3), \theta^1 = (20, 2)$. Their density functions are shown in Figure 4:

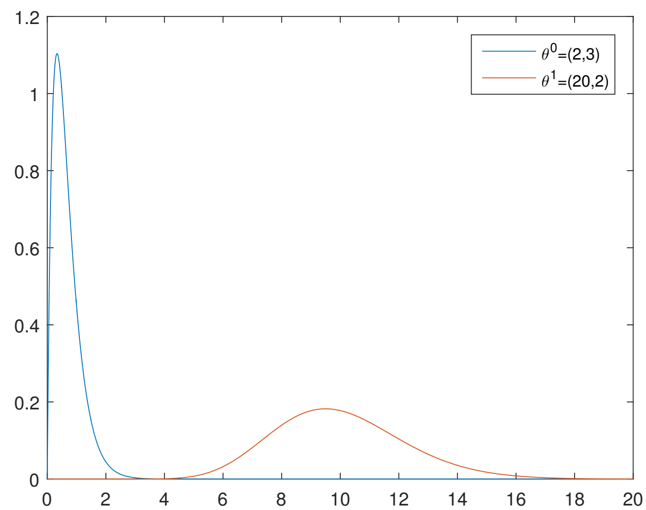

FIGURE 4. Gamma density functions

We compute the related geodesic, in the Wasserstein statistical manifold and the whole density space respectively. The results are presented in Figure 5. We can see that these two do not differ very much. This means the optimal transport in the whole space could nearly keep the gamma distribution shape along the transportation.

Then, we consider the gamma distribution fitting problem. The model is similar to the one in the mixture examples, except that the parameterized family changes. The minimization problem is:

$$\min_\theta \frac{1}{2} \left(W_2(\rho(\cdot; \theta), \frac{1}{N} \sum_{i=1}^N \delta_{x_i}) \right)^2,$$

where $x_i \sim \rho(\cdot, \theta^1)$ and we set $N = 10^3$. The initial guess is $\theta = \theta^0$. Convergence results are presented in Figure 6.

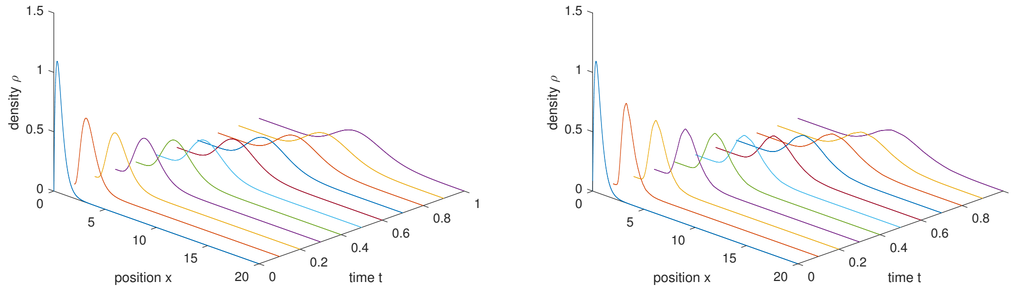

FIGURE 5. Geodesic of Gamma distribution; left: in the Wasserstein statistical manifold; right: in the whole density space

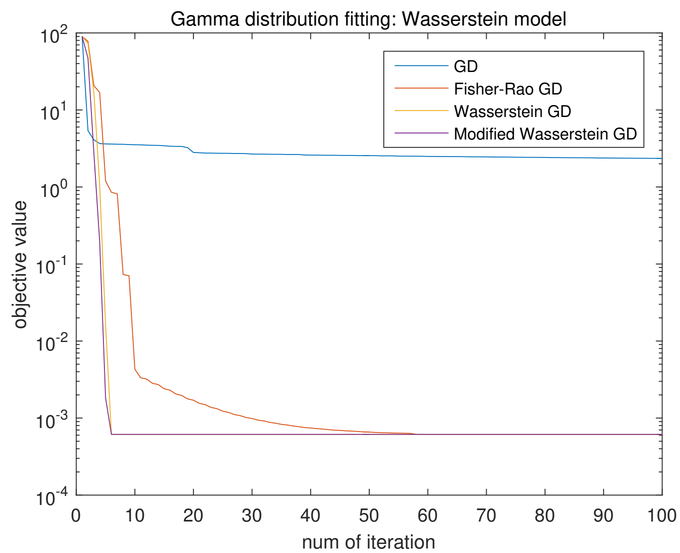

FIGURE 6. objective value

The figure shows that the Euclidean gradient descent method takes very long time to reach convergence, while Wasserstein GD and its modified version needs less than 10 steps, with the Fisher-Rao GD taking around 50 steps. This comparison demonstrates the efficiency of Wasserstein natural gradient in this Wasserstein metric modeled optimization problems. As is mentioned in the previous example, the difference between using G_W and $\bar{G}_W$ is very small. Since $\bar{G}_W$ fails in the mixture example, we conclude that G_W, the Wasserstein gradient descent, will be a more stable choice for preconditioning.

5. Discussion

To summarize, we introduce the Wasserstein statistical manifold for parametric models with continuous sample space. The metric tensor is derived by pulling back the L^2-Wasserstein metric tensor in density space to parameter spaces. Given this Riemannian structure, the

Wasserstein natural gradient is then proposed. In one-dimensional sample space, we obtain an explicit formula for this metric tensor, and from it, we show that the Wasserstein natural gradient descent method achieves asymptotically Newton method for Wasserstein metric modeled minimizations. Our numerical examples justify these arguments.

One potential future direction is using Theorem 1 to design various efficient algorithms for solving Wasserstein metric modeled problems. The Wasserstein gradient descent only takes the asymptotic behavior into consideration, and we think a careful investigation of the structure (13) will lead to better non-asymptotic results. Moreover, generalizing (13) to higher dimensions also remains a challenging and interesting issue. We are working on designing efficient computational method for obtaining $G_W(\theta)$ and hope to report it in subsequent papers.

Analytically, the treatment of the Wasserstein statistical manifold could be generalized. This paper takes an initial step in introducing Wasserstein geometry to parametric models. More analysis on the solution of the elliptic equation and its regularity will be conducted.

Further, we believe ideas and studies from information geometry could lead to natural extensions in Wasserstein statistical manifold. The Wasserstein distance has shown its effectiveness in illustrating and measuring low dimensional supported densities in high dimensional space, which is often the target of many machine learning problems. We are interested in geometric properties of Wasserstein metric in these models, and we will continue to work on it.

Acknowledgments: This Research is partially supported by AFOSR MURI proposal number 18RT0073. The authors thank Prof. Shui-Nee Chow for his farseeing viewpoints on the related topics. We also acknowledge many fruitful discussions with Prof. Wilfrid Gangbo and Prof. Wotao Yin. Part of this work was done when the first author was a visiting undergraduate student in University of California, Los Angeles.

References

[1] S. Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2):251–276, 1998.

[2] S. Amari. *Information Geometry and Its Applications.* Number volume 194 in Applied mathematical sciences. Springer, Japan, 2016.

[3] S. Amari and A. Cichocki. Adaptive blind signal processing-neural network approaches. *Proceedings of the IEEE*, 86(10):2026–2048, 1998.

[4] S. Amari, R. Karakida, and M. Oizumi. Information Geometry Connecting Wasserstein Distance and Kullback-Leibler Divergence via the Entropy-Relaxed Transportation Problem. *arXiv:1709.10219 [cs, math]*, 2017.

[5] L. Ambrosio, N. Gigli, and Savaré Giuseppe. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures.* Birkhäuser Basel, Basel, 2005.

[6] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv:1701.07875 [cs, stat]*, 2017.

[7] N. Ay, J. Jost, H. V. Lê, and L. J. Schwachhöfer. *Information Geometry.* Ergebnisse der Mathematik und ihrer Grenzgebiete A series of modern surveys in mathematics. Folge, volume 64. Springer, Cham, 2017.

[8] J. D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[9] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. Inference in generative models using the wasserstein distance. *arXiv:1701.05146 [math, stat]*, 2017.

[10] E. A. Carlen and W. Gangbo. Constrained Steepest Descent in the 2-Wasserstein Metric. *Annals of Mathematics*, 157(3):807–846, 2003.

[11] F. P. Carli, L. Ning, and T. T. Georgiou. Convex Clustering via Optimal Mass Transport. *arXiv:1307.5459 [cs]*, 2013.

[12] J. Chen, Y. Chen, H. Wu, and D. Yang. The quadratic Wasserstein metric for earthquake location. *arXiv:1710.10447 [math]*, 2017.

[13] Y. Chen, T. T. Georgiou, and A. Tannenbaum. Optimal transport for Gaussian mixture models. *arXiv:1710.07876 [cs, math]*, 2017.

[14] N. N. Chentsov. *Statistical decision rules and optimal inference*. American Mathematical Society, Providence, R.I., 1982.

[15] S. N. Chow, W. Li, J. Lu, and H. Zhou. Population games and Discrete optimal transport. *arXiv:1704.00855 [math]*, 2017.

[16] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, N.J, 2nd ed edition, 2006.

[17] P. Degond, J. G. Liu, and C. Ringhofer. Large-Scale Dynamics of Mean-Field Games Driven by Local Nash Equilibria. *Journal of Nonlinear Science*, 24(1):93–115, 2014.

[18] B. Engquist and B. D. Froese. Application of the Wasserstein metric to seismic signals. *Communications in Mathematical Sciences*, 12(5):979–988, 2014.

[19] B. Engquist, B. D. Froese, and Y. Yang. Optimal transport for seismic full waveform inversion. *Communications in Mathematical Sciences*, 14(8):2309–2330, 2016.

[20] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. Poggio. Learning with a Wasserstein Loss. *arXiv:1506.05439 [cs, stat]*, 2015.

[21] J. D. Lafferty. The density manifold and configuration space quantization. *Transactions of the American Mathematical Society*, 305(2):699–741, 1988.

[22] W. Li. Geometry of probability simplex via optimal transport. *arXiv:1803.06360 [math]*, 2018.

[23] W. Li and G. Montufar. Natural gradient via optimal transport I. *arXiv:1803.07033 [cs, math]*, 2018.

[24] J. Lott. Some Geometric Calculations on Wasserstein Space. *Communications in Mathematical Physics*, 277(2):423–437, 2007.

[25] J. Lott and C. Villani. Ricci curvature for metric-measure spaces via optimal transport. *Annals of Mathematics*, 169(3):903–991, 2009.

[26] L. Malagò, L. Montrucchio, and G. Pistone. Wasserstein Riemannian Geometry of Positive Definite Matrices. *arXiv:1801.09269 [math, stat]*, 2018.

[27] J. Martens. New insights and perspectives on the natural gradient method. *arXiv:1412.1193 [cs, stat]*, 2014.

[28] G. Marti, S. Andler, F. Nielsen, and P. Donnat. Optimal transport vs. Fisher-Rao distance between copulas for clustering multivariate time series. *2016 IEEE Statistical Signal Processing Workshop*, pages 1–5, 2016.

[29] K. Modin. Geometry of Matrix Decompositions Seen Through Optimal Transport and Information Geometry. *Journal of Geometric Mechanics*, 9(3):335–390, 2017.

[30] G. Montavon, K. R. Müller, and M. Cuturi. Wasserstein Training of Restricted Boltzmann Machines. In *Advances in Neural Information Processing Systems 29*, pages 3718–3726. Curran Associates, Inc., 2016.

[31] F. Otto. The geometry of dissipative evolution equations the porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.

[32] G. Peyré and M. Cuturi. Computational Optimal Transport. *arXiv:1803.00567 [stat]*, 2018.

[33] A. Sanctis and S. Gattone. A Comparison between Wasserstein Distance and a Distance Induced by Fisher-Rao Metric in Complex Shapes Clustering. *Proceedings*, 2(4):163, 2017.

[34] S. Shahshahani. *A New Mathematical Framework for the Study of Linkage and Selection*. Number 211 in Memoirs of the American Mathematical Society. Am. Math. Soc, Providence, R.I, 1979.

[35] A. Takatsu. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026, 2011.

[36] C. Villani. *Optimal Transport: Old and New*. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.

[37] T. L. Wong. Logarithmic divergences from optimal transport and renyi geometry. *arXiv:1712.03610 [cs, math, stat]*, 2017.

Department of Mathematical Sciences, Tsinghua University, Beijing, China 100084, email: chenyifan14@mails.tsinghua.edu.cn

Department of Mathematics, UCLA, Los Angeles, CA 90095 USA, email: wcli@math.ucla.edu