NATURAL GRADIENT VIA OPTIMAL TRANSPORT I

WUCHEN LI AND GUIDO MONTÚFAR

ABSTRACT. We study a natural Wasserstein gradient flow on manifolds of probability distributions with discrete sample spaces. We derive the Riemannian structure for the probability simplex from the dynamical formulation of the Wasserstein distance on a weighted graph. We pull back the geometric structure to the parameter space of any given probability model, which allows us to define a natural gradient flow there. In contrast to the natural Fisher-Rao gradient, the natural Wasserstein gradient incorporates a ground metric on sample space. We discuss implementations following the forward and backward Euler methods. We illustrate the analysis on elementary exponential family examples.

1. INTRODUCTION

The statistical distance between histograms plays a fundamental role in statistics and machine learning. It provides the geometric structure on statistical manifolds [3]. Learning problems usually correspond to minimizing a loss function over these manifolds. An important example is the Fisher-Rao metric on the probability simplex, which has been studied especially within the field of information geometry [3, 6]. A classic result due to Chentsov [10] characterizes this Riemannian metric as the only one, up to scaling, that is invariant with respect to natural statistical embeddings by Markov morphisms (see also [8, 18, 28]). Using the Fisher-Rao metric, a natural Riemannian gradient descent method is introduced [2]. This natural gradient has found numerous successful applications in machine learning (see, e.g., [1, 32, 36, 23, 31]).

Optimal transport provides another statistical distance, named Wasserstein or Earth Mover's distance. In recent years, this metric has attracted increasing attention within the machine learning community [5, 15, 27]. One distinct feature of optimal transport is that it provides a distance among histograms that incorporates a ground metric on sample space. In particular, the L^2-Wasserstein distance has a dynamical formulation, which exhibits a metric tensor structure. The set of probability densities with this metric forms an infinite-dimensional Riemannian manifold, named density manifold [17]. The gradient descent method in the density manifold, called Wasserstein gradient flow, has been widely studied in the literature; see [30, 34] and references.

A question intersecting optimal transport and information geometry arises: What is the natural Wasserstein gradient descent method on the parameter space of a statistical model? In optimal transport, the Wasserstein gradient flow is studied on the full space of probability densities, and shown to have deep connections with the ground metrics on

Key words and phrases. Optimal transport; Information geometry; Wasserstein statistical manifold; Machine learning.

sample space deriving from physics [29], fluid mechanics [9] and differential geometry [21]. We expect that these relations also exist on parametrized probability models, and that the Wasserstein gradient flow can be useful in the optimization of objective functions that arise in machine learning problems. By incorporating a ground metric on sample space, this method can serve to implement useful priors in the learning algorithms.

We are interested in developing synergies between the information geometry and optimal transport communities. In this paper we take a natural first step in this direction. We introduce the Wasserstein natural gradient flow on the parameter space of probability models with discrete sample spaces. The L^2-Wasserstein metric on discrete states was introduced in [11, 22, 25]. Following the settings from [12, 13, 16, 19], the probability simplex forms the Riemannian manifold called Wasserstein probability manifold. The Wasserstein metric on the probability simplex can be pulled back to the parameter space of a probability model. This metric allows us to define a natural Wasserstein gradient method on parameter space.

In the literature one finds several formulations of optimal transport for continuous sample spaces. On the one hand, there is the static formulation, known as Kantorovich's linear programming [34]. Here, the linear program is to find the minimal value of a functional over the set of joint measures with given marginal histograms. The objective functional is given as the expectation value of the ground metric with respect to a joint probability density measure. On the other hand, there is the dynamical formulation, known as the Benamou-Brenier formula [7]. This dynamic formulation gives the metric tensor for measures by lifting the ground metric tensor of sample spaces.

Both static and dynamic formulations are equivalent in the case of continuous state spaces. However, the two formulations lead to different metrics in the simplex of discrete probability distributions. The major reason for this difference is that the discrete sample space is not a length space.[1] Thus the equivalence result in classical optimal transport is no longer true in the setting of discrete sample spaces. We note that for the static formulation, it defines the distance for the discrete probability simplex, in which the gradient operator is a metric derivative depending on the dual variable of linear programming. In this setting, there is no Riemannian metric tensor for the discrete probability simplex. See [13, 22] and the Appendix for a more detailed discussion.

The connection between optimal transport and information geometry has been studied in [4, 35]. These works mainly use the static formulation of optimal transport on discrete sample spaces. Many connections between static optimal transport and information geometry have been found. In contrast to these works, we focus on the dynamical formulation of optimal transport in order to define a Riemannian metric tensor on the discrete probability simplex. Other works have studied the Gaussian family of distributions with L^2-Wasserstein metric in continuous sample space [26, 33]. Compared to them, we consider discrete state spaces and arbitrary parametric models.

This paper is organized as follows. In Section 2 we briefly review the Riemannian manifold structure in probability space introduced by optimal transport in the cases of

[1] A length space is one in which the distance between points can be measured as the infimum length of continuous curves between them.

continuous and discrete sample spaces. In Section 3 we introduce Wasserstein statistical manifolds by isometric embedding into the probability manifold, and derive the corresponding gradient flows. In Section 4 we discuss a few examples.

2. Optimal transport on continuous and discrete sample spaces

In this section, we briefly review the results of optimal transport. In particular, we derive the corresponding Riemannian manifold structure for probability distributions with a discrete support set.

2.1. **Optimal transport on continuous sample space.** We start with a review of the optimal transport problem on continuous spaces. This will guide our discussion of the finite state case. For related studies, we refer the reader to [17, 34] and the many references therein.

Denote the sample space by (Ω, g^Ω). Here Ω is a finite dimensional smooth Riemannian manifold, for example $\mathbb{R}^d$ or the open unit ball therein. Its inner product is denoted by g^Ω and its volume form by dx. Denote the geodesic distance of Ω by $d_\Omega\colon \Omega\times\Omega\to\mathbb{R}_+$.

Consider the set $\mathcal{P}_2(\Omega)$ of Borel measurable probability density functions on Ω with finite second moment. Given $\rho^0, \rho^1 \in \mathcal{P}_2(\Omega)$, the L^2-Wasserstein distance (also called Wasserstein metric in the optimal transport literature) between ρ^0 and ρ^1 is denoted by $W\colon \mathcal{P}(\Omega)\times\mathcal{P}(\Omega)\to\mathbb{R}_+$. There are two equivalent ways of defining this distance. On one hand, there is the static formulation. This refers to the following linear programming problem:

$$W(\rho^0,\rho^1)^2 = \inf_{\pi\in\Pi(\rho^0,\rho^1)}\int_{\Omega\times\Omega} d_\Omega(x,y)^2\pi(dx,dy)\ , \tag{1}$$

where the infimum is taken over the set Π of joint probability measures on $\Omega\times\Omega$ that have marginals ρ^0, ρ^1.

On the other hand, the Wasserstein distance W can be written in a dynamical formulation, where a probability path $\rho : [0,1]\to\mathcal{P}_2(\Omega)$ connecting ρ^0, ρ^1 is considered. This refers to a variational problem known as the Benamou-Brenier formula:

$$W(\rho^0,\rho^1)^2 = \inf_{\Phi}\int_0^1\int_\Omega g_x^\Omega(\nabla\Phi(t,x),\nabla\Phi(t,x))\rho(t,x)dxdt\ , \tag{2a}$$

where the infimum is taken over the set of Borel *potential* functions $[0,1]\times\Omega\to\mathbb{R}$. Each potential function Φ determines a corresponding density path ρ as the solution of the *continuity equation*

$$\frac{\partial\rho(t,x)}{\partial t} + \operatorname{div}(\rho(t,x)\nabla\Phi(t,x)) = 0\ ,\quad \rho(0,x)=\rho^0(x)\ ,\quad \rho(1,x)=\rho^1(x)\ . \tag{2b}$$

Here div and ∇ are the divergence and gradient operators in Ω. The continuity equation is well known in physics.

The equivalence of the static (1) and dynamical (2) formulations is well known. For the reader's convenience we give a sketch of proof in the appendix. In this paper we focus on the variational formulation (2). In fact, this formulation entails the definition of a Riemannian structure as we now discuss. For simplicity, we only consider the set of smooth and strictly positive probability 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 $\mathcal{F}(\Omega) := C^\infty(\Omega)$ 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

$$V_\Phi(x) := -\mathrm{div}(\rho(x)\nabla\Phi(x))\ .$$

We assume the zero flux condition

$$\int_\Omega V_\Phi(x)dx = 0\ .$$

In view of the continuity equation, the zero flux condition is equivalent to requiring that $\int_\Omega \frac{\partial\rho}{\partial t}dx = 0$, which means that the space integral of ρ is always 1. When Ω is compact without boundary, this is automatically satisfied. This is also true when $\Omega = \mathbb{R}^d$ and ρ has finite second moment. Thus $V_\Phi \in T_\rho\mathcal{P}_+(\Omega)$. The elliptic operator $\nabla\cdot(\rho\nabla)$ identifies the function Φ on Ω modulo additive constants with the tangent vector V_Φ of the space of densities (for more details see [17, 21]). This gives an isomorphism

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

Define the Riemannian metric (inner product) on the tangent space of positive densities $g^W \colon T_\rho\mathcal{P}_+(\Omega)\times T_\rho\mathcal{P}_+(\Omega) \to \mathbb{R}$ by

$$g^W_\rho(V_\Phi, V_{\tilde\Phi}) = \int_\Omega g^\Omega_x(\nabla\Phi(x), \nabla\tilde\Phi(x))\rho(x)dx\ ,$$

where $\Phi(x),\ \tilde\Phi(x) \in \mathcal{F}(\Omega)/\mathbb{R}$. The inner product endows $\mathcal{P}_+(\Omega)$ with an infinite dimensional Riemannian metric tensor. In other words, the variational problem (2) is a geometric action energy [7, 21] in $(\mathcal{P}_+(\Omega), g^W)$. In literature [17], $(\mathcal{P}_+(\Omega), g^W)$ is called density manifold.

2.2. **Dynamical optimal transport on discrete sample spaces.** We translate the dynamical perspective from the previous section to discrete state spaces, i.e., we replace the continuous space Ω by the discrete space $I = \{1, \cdots, n\}$.

To encode the metric tensor of discrete states, we first need to introduce a ground metric notion on sample space. We do this in terms of a graph with weighted edges, $G = (V, E, \omega)$, where $V = I$ is the vertex set, E is the edge set, and $\omega = (\omega_{ij})_{i,j\in I} \in \mathbb{R}^{n\times n}$ are the edge weights. These weights satisfy

$$\omega_{ij} = \begin{cases}\omega_{ji} > 0, & \text{if } (i,j) \in E\\ 0, & \text{otherwise}\end{cases}\ .$$

As mentioned above, the weights encode the ground metric on the discrete state space. More precisely, we write

$$\omega_{ij} = \frac{1}{(d_{ij}^G)^2} \;, \quad \text{if } (i,j) \in E \;, \tag{3}$$

where d_{ij}^G represents the distance between states i and j. The set of neighbors or adjacent vertices of i is denoted by $N(i) = \{j \in V : (i,j) \in E\}$.

The probability simplex supported on the vertices of G is defined by

$$\mathcal{P}(I) = \Big\{ (p_1, \cdots, p_n) \in \mathbb{R}^n \colon \sum_{i=i}^n p_i = 1 \;, \quad p_i \geq 0 \Big\} \;.$$

Here $p = (p_1, \dots, p_n)$ is a probability vector with coordinates p_i corresponding to the probabilities assigned to each node i. We denote the interior of the probability simplex by $\mathcal{P}_+(I)$. This consists of the strictly positive probability distributions, with $p_i > 0$, $i \in I$.

Next we introduce the variational problem (2) on discrete states. First we need to define the "metric tensor" on graphs. A *vector field* $v = (v_{ij})_{i,j \in V} \in \mathbb{R}^{n \times n}$ on G is a skew-symmetric matrix:

$$v_{ij} = \begin{cases} -v_{ji}, & \text{if } (i,j) \in E \\ 0, & \text{otherwise} \end{cases} \;.$$

A potential function $\Phi = (\Phi_i)_{i=1}^n \in \mathbb{R}^n$ defines *gradient vector field* $\nabla_G \Phi = (\nabla_G \Phi_{ij})_{i,j \in V} \in \mathbb{R}^{n \times n}$ on the graph G by the finite differences

$$\nabla_G \Phi_{ij} = \begin{cases} \sqrt{\omega_{ij}}(\Phi_i - \Phi_j) & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases} \;.$$

Here we use $\sqrt{\omega}$ rather than the more natural $1/d$. This is for the simplicity of notations. So that we can represent the gradient, divergence and Laplacian matrix in a multiplicity of weight, instead of dividing the ground metric.

We define an inner product of tangent vectors at each state $i \in I$ by

$$g_i^I(v,v) := \frac{1}{2} \sum_{j \in N(i)} v_{ij}^2 \;.$$

In particular, the gradient vector field $\nabla_G \Phi$ defines a kinetic energy at each state $i \in I$ by

$$g_i^I(\nabla_G \Phi, \nabla_G \Phi) := \frac{1}{2} \sum_{j \in N(i)} (\Phi_i - \Phi_j)^2 \omega_{ij}$$

We next define the expectation value of kinetic energy with respect to the probability distribution p:

$$(\nabla_G \Phi, \nabla_G \Phi)_p := \sum_{i \in I} p_i \; g_i^I(\nabla_G \Phi, \nabla_G \Phi) = \frac{1}{2} \sum_{(i,j) \in E} \omega_{ij} (\Phi_i - \Phi_j)^2 \frac{p_i + p_j}{2} \;.$$

The latter can also be written as

$$(\nabla_G \Phi, \nabla_G \Phi)_p = \sum_{i=1}^n \Phi_i \sum_{j \in N(i)} \omega_{ij} (\Phi_i - \Phi_j) \frac{p_i + p_j}{2} = \Phi^{\mathsf{T}} \big(-\operatorname{div}_G(p \nabla_G \Phi) \big) \;,$$

where

$$-\operatorname{div}_G(p\nabla_G\Phi) := \Big(\sum_{j\in N(i)} \omega_{ij}(\Phi_i - \Phi_j)\frac{p_i + p_j}{2}\Big)_{i\in I} . \tag{4}$$

There are two definitions hidden in (4). First, $\operatorname{div}_G\colon \mathbb{R}^{n\times n} \to \mathbb{R}^n$ maps any given vector field m on the graph to

$$\operatorname{div}_G(m) = (\sum_{j\in N(i)} \sqrt{\omega_{ij}} m_{ji})_{i=1}^n .$$

Second, the probability weight $m = p\nabla_G\Phi$ of the gradient vector field $\nabla_G\Phi$ is defined by

$$m_{ij} = \begin{cases} \frac{p_i+p_j}{2}(\Phi_i - \Phi_j)\sqrt{\omega_{ij}}, & \text{if } (i,j)\in E \\ 0, & \text{otherwise} \end{cases} ,$$

where $\frac{p_i+p_j}{2}$ represents the probability weight on the edge $(i,j)\in E$.

Now we are ready to introduce the L^2-Wasserstein metric on $\mathcal{P}_+(I)$.

Definition 1. *For any p^0, $p^1 \in \mathcal{P}_+(I)$, define the Wasserstein distance $W\colon \mathcal{P}_+(I)\times\mathcal{P}_+(I)\to\mathbb{R}$ by*

$$W(p^0,p^1)^2 := \inf_{p(t),\Phi(t)} \Big\{ \int_0^1 (\nabla_G\Phi(t), \nabla_G\Phi(t))_{p(t)} dt \Big\} .$$

Here the infimum is taken over pairs $(p(t),\Phi(t))$ with $p\in H^1((0,1),\mathbb{R}^n)$ and $\Phi\colon [0,1]\to\mathbb{R}^n$ measurable, satisfying

$$\frac{d}{dt}p(t) + \operatorname{div}_G(p(t)\nabla_G\Phi(t)) = 0,\ p(0) = p^0,\ p(1) = p^1.$$

Remark 1. It is worth mentioning that the metric given in Definition 1 is different from the metric defined by linear programming. In other words, denote the distance $d^G(i,j)$ between two vertices i and j is the length of a shortest (i,j)-path. If $(i,j)\in E$, then $d^G(i,j)$ is same as the ground metric defined in (3). Then

$$\big(W(p^0,p^1)\big)^2 \not\equiv \min_\pi \Big\{ \sum_{1\le i,j\le n} d_G(i,j)^2\pi_{ij} \ :\ \sum_{i=1}^n \pi_{ij} = p_j^0 ,\quad \sum_{j=1}^n \pi_{ij} = p_i^1 ,\quad \pi_{ij}\ge 0\Big\} . \tag{5}$$

The reason for this in-equivalence is that discrete sample space I is not a length space. In other words, there is no continuous path in I connecting two nodes in I. For more details see discussions in the appendix.

2.3. **Wasserstein manifold of discrete probability distributions.** In this section we introduce the primal coordinates of the discrete probability simplex with L^2 Wasserstein Riemannian metric. Our discussion follows the recent work [19]. The probability simplex $\mathcal{P}(I)$ is a manifold with boundary. To simplify the discussion, we focus on the interior $\mathcal{P}_+(I)$. The geodesic properties on the boundary $\partial\mathcal{P}(I)$ have been studied in [16].

Let us focus on the Riemannian structure. In the following we introduce an inner product on the tangent space

$$T_p\mathcal{P}_+(I) = \Big\{(\sigma_i)_{i=1}^n \in \mathbb{R}^n\colon \sum_{i=1}^n \sigma_i = 0\Big\} .$$

Denote the space of potential function on I by $\mathcal{F}(I) = \mathbb{R}^n$. Consider the quotient space

$$\mathcal{F}(I)/\mathbb{R} = \{[\Phi] \mid (\Phi_i)_{i=1}^n \in \mathbb{R}^n\} \ ,$$

where $[\Phi] = \{(\Phi_1 + c, \cdots, \Phi_n + c) \mid c \in \mathbb{R}\}$ are functions defined up to addition of constants.

We introduce an identification map via (4)

$$V \colon \mathcal{F}(I)/\mathbb{R} \to T_p\mathcal{P}_+(I) \ , \qquad V_\Phi = -\mathrm{div}_G(p\nabla_G\Phi) \ .$$

In [11] it is shown that $V_\Phi \colon \mathcal{F}(I)/\mathbb{R} \to T_p\mathcal{P}_+(I)$ is a well defined map, linear, and one to one. I.e., $\mathcal{F}(I)/\mathbb{R} \cong T_p^*\mathcal{P}_+(I)$, where $T_p^*\mathcal{P}_+(I)$ is the cotangent space of $\mathcal{P}_+(I)$. This identification induces the following inner product on $T_p\mathcal{P}_+(I)$.

We first present this in a *dual* formulation, which is known in the literature [21].

Definition 2 (Inner product in dual coordinates)**.** *The inner product $g_p^W : T_p\mathcal{P}_+(I) \times T_p\mathcal{P}_+(I) \to \mathbb{R}$ takes any two tangent vectors $\sigma^1 = V_{\Phi_1}$ and $\sigma^2 = V_{\Phi_2} \in T_p\mathcal{P}_+(I)$ to*

$$g_p^W(\sigma_1, \sigma_2) = \sigma_1^\mathsf{T}\Phi_2 = \sigma_2^\mathsf{T}\Phi_1 = (\nabla_G\Phi_1, \nabla_G\Phi_2)_p \ . \tag{6}$$

The above is written in the dual coordinates. We shall now give the inner product in primal coordinates. The following matrix operator will be the key to the Riemannian metric tensor of $(\mathcal{P}_+(I), g^W)$.

Definition 3 (Linear weighted Laplacian matrix)**.** *Given a weighted graph $G = (I, E, \omega)$, $I = \{1, \dots, n\}$, the matrix function $L(\cdot) : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is defined by*

$$L(a) = D^\mathsf{T}\Lambda(a)D \ , \quad a = (a_i)_{i=1}^n \in \mathbb{R}^n \ ,$$

where

- *$D \in \mathbb{R}^{|E|\times n}$ is the discrete gradient operator*

$$D_{(i,j)\in E, k\in V} = \begin{cases} \sqrt{\omega_{ij}}, & \text{if } i = k,\ i > j \\ -\sqrt{\omega_{ij}}, & \text{if } j = k,\ i > j \\ 0, & \text{otherwise} \end{cases} \ ,$$

- *$-D^\mathsf{T} \in \mathbb{R}^{n\times|E|}$ is the discrete divergence operator, also named oriented incidence matrix* [14]*, and*
- *$\Lambda(a) \in \mathbb{R}^{|E|\times|E|}$ is a weight matrix depending on a,*

$$\Lambda(a)_{(i,j)\in E, (k,l)\in E} = \begin{cases} \frac{a_i + a_j}{2} & \text{if } (i,j) = (k,l) \in E \\ 0 & \text{otherwise} \end{cases} \ .$$

Consider some $p \in \mathcal{P}_+(I)$. From spectral graph theory [14], we know that $L(p)$ can be decomposed as

$$L(p) = U(p) \begin{pmatrix} 0 & & & \\ & \lambda_1(p) & & \\ & & \ddots & \\ & & & \lambda_{n-1}(p) \end{pmatrix} U(p)^\mathsf{T} \ .$$

Here $0 < \lambda_1(p) \leq \cdots \leq \lambda_{n-1}(p)$ are the eigenvalues of $L(p)$ in ascending order, and $U(p) = (u_0(p), u_1(p), \cdots, u_{n-1}(p))$ is the corresponding orthogonal matrix of eigenvectors with

$$u_0 = \frac{1}{\sqrt{n}}(1, \cdots, 1)^\mathsf{T} .$$

We write $L(p)^\dagger$ for the pseudo-inverse of $L(p)$, i.e.,

$$L(p)^\dagger = U(p) \begin{pmatrix} 0 & & & \\ & \frac{1}{\lambda_1(p)} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_{n-1}(p)} \end{pmatrix} U(p)^\mathsf{T} .$$

With $\sigma_i = L(p)\Phi_i$, $i = 1, 2$, we see that

$$\sigma_1{}^\mathsf{T} L(p)^\dagger \sigma_2 = \Phi_1^\mathsf{T} L(p) L(p)^\dagger L(p) \Phi_2 = \Phi_1^\mathsf{T} L(p) \Phi_2 = (\nabla_G \Phi_1, \nabla_G \Phi_2)_p .$$

Now we are ready to give the inner product in primal coordinates.

Definition 4 (Inner product in primal coordinates)**.** *The inner product* $g_p^W : T_p\mathcal{P}_+(I) \times T_p\mathcal{P}_+(I) \to \mathbb{R}$ *is defined by*

$$g_p^W(\sigma_1, \sigma_2) := \sigma_1{}^\mathsf{T} L(p)^\dagger \sigma_2 , \quad \textit{for any } \sigma_1, \sigma_2 \in T_p\mathcal{P}_+(I) .$$

In other words, the variational problem from Definition 1 is a minimization of geometry energy functional in $\mathcal{P}_+(I)$, i.e.,

$$W(p^0, p^1)^2 = \inf_{p(t) \in \mathcal{P}_+(I)} \left\{ \int_0^1 \dot{p}(t)^\mathsf{T} L(p(t))^\dagger \dot{p}(t) dt \, : \, p(0) = p^0 \ , \ p(1) = p^1 \right\} .$$

This defines a Wasserstein Riemannian strucure on the probability simplex. For more details of Riemannian formulas see [19]. Following [17] we could call $(\mathcal{P}_+(I), g^W)$ discrete density manifold. However, this could be easily confused with other notions from information geometry, and hence we will use the more explicit terminology *Wasserstein discrete probability manifold*, or Wasserstein manifold for short.

3. Natural gradient for the Wasserstein manifold

In this section we study parametric probability models endowed with the L^2-Wasserstein Riemannian metric. We define this in the natural way, by pulling back the Riemannian structure from the Wasserstein manifold that we discussed in the previous section. This allows us to introduce a natural gradient flow on the parameter space of statistical models.

3.1. **Wasserstein statistical manifold.** Consider a statistical model defined by a triplet $(\Theta, I, \mathbf{p})$. Here, $I = \{1, \cdots, n\}$ is the sample space, Θ is the parameter space, which is an open subset of $\mathbb{R}^d$, $d \leq n - 1$, and $\mathbf{p} \colon \Theta \to \mathcal{P}_+(I)$ is the parametrization function,

$$\mathbf{p}(\xi) = (p_i(\xi))_{i=1}^n , \quad \xi \in \Theta .$$

In the sequel we will assume that $\operatorname{rank}(J_\xi \mathbf{p}(\xi)) = d$, so that the parametrization is locally injective.

We define a Riemannian metric g on Θ as the pull-back of metric g^W on $\mathcal{P}_+(I)$. In other words, we require that $\mathbf{p}\colon (\Theta, g) \to (\mathcal{P}_+(I), g^W)$ is an isometric embedding:

$$\begin{aligned} g_\xi(a,b) :=& g^W_{\mathbf{p}(\xi)}(d\mathbf{p}(\xi)(a), d\mathbf{p}(\xi)(b)) \\ =& \big(d\mathbf{p}(\xi)(a)\big)^\mathsf{T} L(\mathbf{p}(\xi))^\dagger \big(d\mathbf{p}(\xi)(b)\big) \ . \end{aligned}$$

Since $d\mathbf{p}(\xi)(a) = \big(\sum_{j=1}^n \frac{\partial p_i(\xi)}{\partial \xi_j} a_j\big)_{i=1}^n = J_\xi \mathbf{p}(\xi) a$, we arrive at the following definition.

Definition 5. *For any pair of tangent vectors $a, b \in T_\xi\Theta = \mathbb{R}^d$, define*

$$g_\xi(a,b) := a^\mathsf{T} \big(J_\xi \mathbf{p}(\xi)\big)^\mathsf{T} L(\mathbf{p}(\xi))^\dagger (J_\xi \mathbf{p}(\xi)) b \ ,$$

where $J_\xi(\mathbf{p}(\xi)) = (\frac{\partial p_i(\xi)}{\partial \xi_j})_{1\le i\le n, 1\le j\le d} \in \mathbb{R}^{n\times d}$ is the Jacobi matrix of the parametrization $\mathbf{p}$, and $L(\mathbf{p}(\xi)) \in \mathbb{R}^{n\times n}$ is the pseudo-inverse of the linear weighted Laplacian matrix.

This inner product is consistent with the restriction of the Wasserstein metric g^W to $\mathbf{p}(\Theta)$. For this reason, we call $\mathbf{p}(\Theta)$, or $(\Theta, I, \mathbf{p})$, together with the induced Riemannian metric g, *Wasserstein statistical manifold.*

We need to make sure that the embedding procedure is valid, because the metric tensor $L(\mathbf{p})^\dagger$ is only of rank $n-1$. The next lemma shows that (Θ, g) is a well defined d-dimensional Riemannian manifold.

Lemma 6. *For any $\xi \in \Theta$, we have*

$$\lambda_{\min}(\xi) = \inf_{a\in\mathbb{R}^d, \|a\|_2=1} g_\xi(a,a) > 0 \ .$$

In addition, g_ξ is smooth as a function of ξ, so that (Θ, g) is a smooth Riemannian manifold.

Proof. We only need to show that $\big(J_\xi \mathbf{p}(\xi)\big)^\mathsf{T} L(\mathbf{p}(\xi))^\dagger \big(J_\xi \mathbf{p}(\xi)\big) \in \mathbb{R}^{d\times d}$ is a positive definite matrix. Consider

$$a^\mathsf{T}\big(J_\xi \mathbf{p}(\xi)\big)^\mathsf{T} L(\mathbf{p}(\xi))^\dagger \big(J_\xi \mathbf{p}(\xi)\big) a = 0 \ ,$$

where $0 \in \mathbb{R}^{n-1}$. Since $L(p)$ only has one simple eigenvalue 0 with eigenvector u_0, then

$$J_\xi \mathbf{p}(\xi) a = c u_0 \ , \quad \text{for some constant } c \in \mathbb{R}^1 \ . \tag{7}$$

Since $u_0^\mathsf{T} \mathbf{p}(\xi) = \frac{1}{\sqrt{n}} \sum_{i=1}^n p_i(\xi) = 0$, we have that $u_0^\mathsf{T} \frac{\partial \mathbf{p}(\xi)}{\partial \xi_j} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial p_i(\xi)}{\partial \xi_j} = 0$, i.e.,

$$u_0^\mathsf{T} J_\xi \mathbf{p}(\xi) = 0 \ .$$

Left multiply u_0 into (11), we obtain

$$0 = u_0^\mathsf{T} J_\xi \mathbf{p}(\xi) a = c u_0^\mathsf{T} u_0 = c \ .$$

Thus $c = 0$, and (11) forms

$$J_\xi \mathbf{p}(\xi) a = 0 \ .$$

Since $\operatorname{rank}(J_\xi \mathbf{p}(\xi)) = d < n$, we have $a = 0$, which finishes the proof. □

We briefly illustrate some geometric calculations on (Θ, g). Given $\xi_0,\ \xi_1 \in \Theta$, the Riemannian distance on (Θ, g) is given by

$$\mathrm{Dist}(\xi_0,\xi_1)^2 = \inf_{\xi(t)\in C^1((0,1);\Theta)} \left\{ \int_0^1 \dot{\xi}^{\mathsf{T}} G(\xi)\dot{\xi} \ :\ \xi(0)=\xi_0,\ \xi(1)=\xi_1 \right\} .$$

where the metric tensor is given by

$$G(\xi) = (J_\xi p(\xi))^{\mathsf{T}} L(p(\xi))^{\dagger} (J_\xi p(\xi)) \ . \tag{8}$$

Following the standard Riemannian geometry, the cotangent geodesic flow forms

$$\begin{cases} \dot{\xi} - G(\xi)^{-1} S = 0 \\ \dot{S} + \frac{1}{2}\frac{\partial}{\partial \xi} S^T G(\xi)^{-1} S = 0 \end{cases} . \tag{9}$$

Here S is the Legendre transformation of $\dot{\xi}$. It is worth recalling the following facts. If $\mathbf{p}$ is an identity map, then (9) translates to

$$\begin{cases} \dot{p} + \mathrm{div}_G(p\nabla_G S) = 0 \\ \dot{S} + \frac{1}{4}\sum_{j\in N(i)}(\nabla_G S)^2 = 0 \end{cases} .$$

In addition, if $I=\Omega$ and we replace i by x and $p_i(t)$ by $\rho(t,x)$, the above becomes

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

which are the standard continuity and Hamilton-Jacobi equations on Ω. For these reasons, we call the two equations in (9) the *continuity equation* and the *Hamilton-Jacobi equation in parameter space.*

3.2. Gradient flow on Wasserstein statistical manifold.

In this section we introduce the natural Riemannian gradient flow on (Θ, g).

Consider a loss function $F\colon \mathcal{P}_+(I) \to \mathbb{R}$. Thus we focus on the composition $F\circ \mathbf{p}\colon \Theta \to \mathbb{R}$. The Riemannian gradient of $F(\mathbf{p}(\xi))$ is defined as follows. Given some $\nabla_g F(\mathbf{p}(\xi)) \in T_\xi\Theta$, we have

$$g_\xi(\nabla_g F(\mathbf{p}(\xi)), a) = \mathrm{d}_\xi F(\mathbf{p}(\xi))\cdot a\ , \quad \text{for any } a\in T_\xi\Theta\ , \tag{10}$$

where $\mathrm{d}_\xi F(\mathbf{p}(\xi))\cdot a = \sum_{i=1}^d \frac{\partial}{\partial \xi_i} F(\mathbf{p}(\xi)) a_i$. The gradient flow satisfies

$$\frac{d\xi}{dt} = -\nabla_g F(\mathbf{p}(\xi))\ .$$

The next theorem establishes an explicit formulation of the gradient flow.

Theorem 7 (Wasserstein gradient flow)**.** *The gradient flow of a functional $F\colon \mathcal{P}_+(I)\to\mathbb{R}$ is given by*

$$\frac{d\xi}{dt} = -G(\xi)^{-1}\nabla_\xi F(\mathbf{p}(\xi))\ ,$$

where ∇_ξ is the Euclidean gradient of $F(\mathbf{p}(\xi))$ w.r.t. ξ. More explicitly,

$$\frac{d\xi}{dt} = -\Big((J_\xi\mathbf{p}(\xi))^{\mathsf{T}} L(\mathbf{p}(\xi))^{\dagger} J_\xi\mathbf{p}(\xi)\Big)^{\dagger} (J_\xi\mathbf{p}(\xi))^{\mathsf{T}} \nabla_{\mathbf{p}} F(\mathbf{p}(\xi)\ , \tag{11}$$

where $\nabla_{\mathbf{p}}$ is the Euclidean gradient of $F(\mathbf{p})$ w.r.t. $\mathbf{p}$.

Proof. The proof follows directly from (10). Notice that

$$g_\xi(\nabla_g F(\mathbf{p}(\xi)), a) = \nabla_g F(\mathbf{p}(\xi))^\mathsf{T} \big(J_\xi \mathbf{p}(\xi)\big)^\mathsf{T} L(\mathbf{p}(\xi))^\dagger (J_\xi \mathbf{p}(\xi)) a = \nabla_\xi F(\mathbf{p}(\xi))^\mathsf{T} a \ ,$$

and $\big(J_\xi \mathbf{p}(\xi)\big)^\mathsf{T} L(\mathbf{p}(\xi))^\dagger (J_\xi \mathbf{p}(\xi))$ is an invertible matrix. Hence

$$\nabla_g F(\mathbf{p}(\xi)) = \big((J_\xi \mathbf{p}(\xi))^\mathsf{T} L(\mathbf{p}(\xi))^\dagger J_\xi \mathbf{p}(\xi)\big)^\dagger d_\xi F(\mathbf{p}(\xi)) \ .$$

We compute $d_\xi F(\mathbf{p}(\xi))$ as

$$d_\xi F(\mathbf{p}(\xi)) = \big(\frac{\partial}{\partial \xi_i} F(\mathbf{p}(\xi))\big)_{i=1}^n = \big(\sum_{j=1}^n \frac{\partial}{\partial p_j} F(\mathbf{p}(\xi)) \cdot \frac{\partial p_j(\xi)}{\partial \xi_i}\big)_{i=1}^n = (J_\xi \mathbf{p}(\xi))^\mathsf{T} \nabla_\mathbf{p} F(\mathbf{p}(\xi)) \ .$$

This concludes the proof of (11). □

Equation (11) is the generalization of the Wasserstein gradient flow the probability simplex to the Wasserstein gradient flow on parameter space. If $\mathbf{p}$ is an identity map $\mathbf{p}(\xi) = \xi$, $\xi = (p_i)_{i=1}^n$, with the parameter space Θ equal to the entire probability simplex, then (11) is

$$\frac{dp}{dt} = -L(p)\nabla_p F(p) = \mathrm{div}_G(p \nabla_G d_p F(p)) \ ,$$

which is the Wasserstein gradient flow on the discrete probability simplex. In particular, if $I = \Omega$, then it represents

$$\frac{\partial \rho(t,x)}{\partial t} = \mathrm{div}_x(\rho(t,x)\nabla_x d_\rho \mathcal{F}(\rho)) \ ,$$

which is the Wasserstein gradient flow on Ω. From now on, we call (11) the *Wasserstein gradient flow in parameter space*.

In fact, the definition of Wasserstein gradient flow shares many similarities with the steepest gradient descent defined as follows. Consider

$$\arg\min_{\Delta\xi} F(\mathbf{p}(\xi + \Delta\xi)) \quad \text{s.t.} \quad \frac{1}{2} W(\mathbf{p}(\xi), \mathbf{p}(\xi+\Delta\xi))^2 = \epsilon \ . \tag{12}$$

By taking the second-order Taylor approximation of the Wasserstein distance at ξ, we get

$$W(\mathbf{p}(\xi), \mathbf{p}(\xi+\Delta\xi))^2 = \Delta\xi^\mathsf{T} G(\xi) \Delta\xi + o(\Delta\xi^2) \ ,$$

where $G(\xi)$ is the metric tensor of (Θ, g) defined in (8), inherited from Wasserstein manifold. We take the first-order approximation of $F(\mathbf{p}(\xi+\Delta\xi))$ in (12) by

$$\arg\min_{\Delta\xi} F(\mathbf{p}(\xi)) + \Delta\xi^\mathsf{T} \nabla_\xi F(\mathbf{p}(\xi)) \quad \text{s.t.} \quad \frac{1}{2}\Delta\xi^\mathsf{T} G(\xi) \Delta\xi = \epsilon \ .$$

By the Lagrangian method with Lagrange multiplier λ, we have

$$\Delta\xi = \lambda G(\xi)^{-1} \nabla_\xi F(\mathbf{p}(\xi)) \ .$$

The above derivations lead to the Wasserstein natural gradient

$$\nabla_g F(\mathbf{p}(\xi)) = G(\xi)^{-1} \nabla_\xi F(\mathbf{p}(\xi)) \ .$$

Remark 2. In the standard natural gradient [2], we replace (12) by

$$\arg\min_{\Delta\xi} F(\mathbf{p}(\xi+\Delta\xi)) \quad \text{s.t.} \quad \mathrm{KL}(\mathbf{p}(\xi)\|\mathbf{p}(\xi+\Delta\xi)) = \epsilon \ ,$$

where KL represents the Kullback-Leibler divergence (relative entropy) from $\mathbf{p}(\xi)$ to $\mathbf{p}(\xi+\Delta\xi)$. Our definition just changes the KL-divergence by the Wasserstein metric on a specified graph.

3.3. **Numerical methods.** Given the gradient flow (11), there are two standard choices of time discretization, namely the forward Euler scheme and the backward Euler scheme. Denote the step size by $h > 0$. The forward Euler method of (11) computes a discretized trajectory by

$$\xi^{k+1} = \xi^k - h\nabla_g F(\mathbf{p}(\xi^k)) \ ,$$

while the backward Euler method computes

$$\xi^{k+1} = \arg\min_{\xi\in\Theta} F(\mathbf{p}(\xi)) + \frac{\mathrm{Dist}(\xi,\xi^k)^2}{2h} \ ,$$

where Dist is the geodesic distance in parameter space (Θ, g).

In the information geometry literature, the forward Euler method is often referred to as natural gradient method. In Wasserstein geometry, the backward Euler method in often called the Jordan-Kinderlehrer-Otto (JKO) scheme. In the following we give pseudo code for both numerical methods.

Natural Wasserstein gradient method

for $k = 1, 2, \dots$ while not converged

1. Choose a suitable step size $h_k > 0$;
2. $\xi^{k+1} = \xi^k - h_k\big((J_\xi\mathbf{p}(\xi^k))^\top L(\mathbf{p}(\xi^k))^\dagger J_\xi\mathbf{p}(\xi^k)\big)^\dagger (J_\xi\mathbf{p}(\xi^k))^\top \nabla_{\mathbf{p}} F(\mathbf{p}(\xi^k))$;

end

Natural Jordan-Kinderlehrer-Otto scheme

for $k = 1, 2, \dots$ while not converged

1. Choose a suitable adaptive step size $h_k > 0$;
2. $\xi^{k+1} = \arg\min_{\xi\in\Theta} F(\mathbf{p}(\xi)) + \frac{\mathrm{Dist}(\xi,\xi^k)^2}{2h_k}$;

end

In practice, the forward Euler method is usually easier to implement than the backward Euler method. We would also suggest to implement the natural Wasserstein gradient using this method for minimization problems in machine learning.

As known in optimization, the JKO scheme can also be useful for non-smooth objective functions. Moreover, the backward Euler method is usually unconditionally stable, which means that one can choose a large step size h for computations.

4. Examples

In this section, we discuss the Wasserstein metric and gradient flow for two exponential families of probabilities.

Example 1 (Wasserstein geodesics). *Consider the sample space $I = \{1, 2, 3\}$ with an unweighted graph $1 - 2 - 3$. The probability simplex for this sample space is a triangle in $\mathbb{R}^3$:*

$$\mathcal{P}(I) = \left\{ (p_i)_{i=1}^3 \in \mathbb{R}^3 \ : \ \sum_{i=1}^3 p_i = 1, \quad p_i \geq 0 \right\} .$$

Following Definition 1, the L^2-Wasserstein distance is given by

$$W(p^0, p^1)^2 := \inf_{\Phi(t)} \frac{1}{2} \int_0^1 (\Phi_1(t) - \Phi_2(t))^2 \frac{p_1(t) + p_2(t)}{2} + (\Phi_2(t) - \Phi_3(t))^2 \frac{p_2(t) + p_3(t)}{2} dt, \tag{13}$$

where the infimum is taken over paths $\Phi\colon [0,1] \to \mathbb{R}^3$. Each Φ defines $p\colon [0,1] \to \mathbb{R}^3$ as the solution of the differential equation

$$\begin{cases} \dot{p}_1 = & (\Phi_1 - \Phi_2) \frac{p_1 + p_2}{2} \\ \dot{p}_2 = & (\Phi_2 - \Phi_1) \frac{p_1 + p_2}{2} + (\Phi_2 - \Phi_3) \frac{p_2 + p_3}{2} \\ \dot{p}_3 = & (\Phi_3 - \Phi_2) \frac{p_2 + p_3}{2} \end{cases}$$

with boundary condition $p(0) = p^0$, $p(1) = p^1$.

Consider local coordinates in (13)*. We parametrize a probability vector as $p = (p_1, 1 - p_1 - p_3, p_3)$, with parameters (p_1, p_3). Then* (13) *can be written as*

$$W(p^0, p^1)^2 := \inf_{p(t)\colon p(0)=p^0, \ p(1)=p^1} \left\{ \int_0^1 \frac{\dot{p}_1(t)^2}{1 - p_3(t)} + \frac{\dot{p}_3(t)^2}{1 - p_1(t)} dt \right\} . \tag{14}$$

This can be solved numerically[2] *for different choices of the boundary conditions.*

We fix three distributions

$$q^1 = \frac{1}{8}(6, 1, 1), \quad q^2 = \frac{1}{8}(1, 6, 1), \quad q^3 = \frac{1}{8}(1, 1, 6) \tag{15}$$

and solve (14) *for three choices of the boundary conditions:*

$$p^0 = q^1, p^1 = q^2; \quad p^0 = q^1, p^1 = q^3; \quad p^0 = q^2, p^1 = q^3. \tag{16}$$

This gives us a geodesic triangle between q^1, q^2, q^3, which is illustrated in Figure 1. It can be seen that $(\mathcal{P}_+(G), W)$ has a non Euclidean geometry. Moreover, we see that the geodesics depend on the graph structure on sample space, where state 2 *is qualitatively differently from states* 1 *and* 3*.*

---

[2] We use the *direct method*, which is a standard technique in optimal control. here the time is discretized, and the sum replacing the integral is minimized by means of gradient descent with respect to $(p(t)_i)_{i=1,3,t\in\{t_1,\ldots,t_N\}} \in \mathbb{R}^{2\times N}$. A standard reference for these techniques is [20].

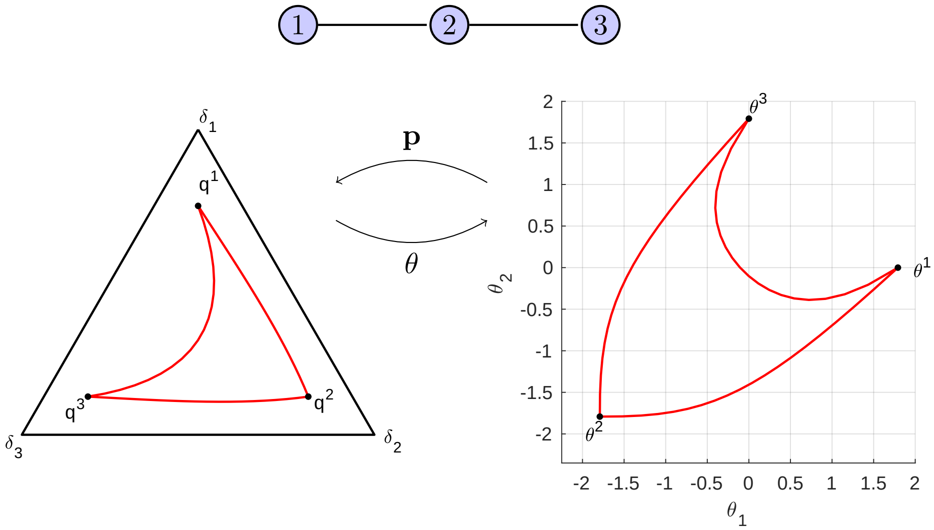

FIGURE 1. The Wasserstein geodesic triangle from Example 1 plotted in the probability simplex (left) and in exponential parameter space (right). The path connecting q^1 and q^3 bends towards q^2; something that does not happen for the other two paths. This illustrates how, as a result of the ground metric on sample space, state 2 is treated differently from 1 and 3.

We can make the same derivations in terms of an exponential parametrization. Consider the parameter space $\Theta = \{\theta = (\theta_1, \theta_2) \in \mathbb{R}^2\}$ *and the parametrization* $\mathbf{p}\colon \Theta \to \mathcal{P}_+(I)$ *with*

$$\mathbf{p}_1(\theta) = \frac{e^{\theta_1}}{e^{\theta_1} + e^{\theta_2} + 1}\,, \quad \mathbf{p}_3(\theta) = \frac{e^{\theta_2}}{e^{\theta_1} + e^{\theta_2} + 1}\,, \quad \mathbf{p}_2(\theta) = 1 - p_1(\theta) - p_3(\theta) = \frac{1}{e^{\theta_1} + e^{\theta_2} + 1}\,.$$

We rewrite the Wasserstein metric (14) *in terms of* θ*. Denote* $\mathbf{p}(\theta^k) = p^k$*,* $k = 0, 1$*. Then the Wasserstein metric in the coordinate system* θ *is*

$$\mathrm{Dist}(\theta^0, \theta^1)^2 = \inf_{\theta(t)\colon \theta(0)=\theta^0,\ \theta(1)=\theta^1} \left\{ \int_0^1 \dot\theta^\intercal J_\theta(\mathbf{p}_1, \mathbf{p}_3)^\intercal \begin{pmatrix} \frac{1}{1-\mathbf{p}_3(\theta)} & 0 \\ 0 & \frac{1}{1-\mathbf{p}_1(\theta)} \end{pmatrix} J_\theta(\mathbf{p}_1, \mathbf{p}_3)\dot\theta dt \right\}.$$

The resulting geodesic triangle in Θ *is plotted in the right panel of Figure 1.*

For comparison, we compute the exponential *geodesic triangle between the same distributions* q^1, q^2, q^3*. This is shown in Figure 2. In this case, there is no distinction between the states* $1, 2, 3$ *and the three paths are symmetric. The exponential geodesic between two distributions* p^0 *and* p^1 *is given by* $(p^0)^{1-t}(p^1)^t / \sum_x (p^0)^{1-t}(p^1)^t$*,* $t \in [0, 1]$*.*

Example 2 (Wasserstein gradient flow on an independence model)**.** *We next illustrate the Wasserstein gradient flow over the independence model of two binary variables. The sample space is* $I = \{-1, +1\}^2$*. For simplicity, we denote the states by* $a = (-1, -1)$*,* $b = (-1, +1)$*,* $c = (+1, -1)$*,* $d = (+1, +1)$*. We consider the square graph*

$$\begin{array}{ccc} b & - & d \\ | & & | \\ a & - & c \end{array}$$

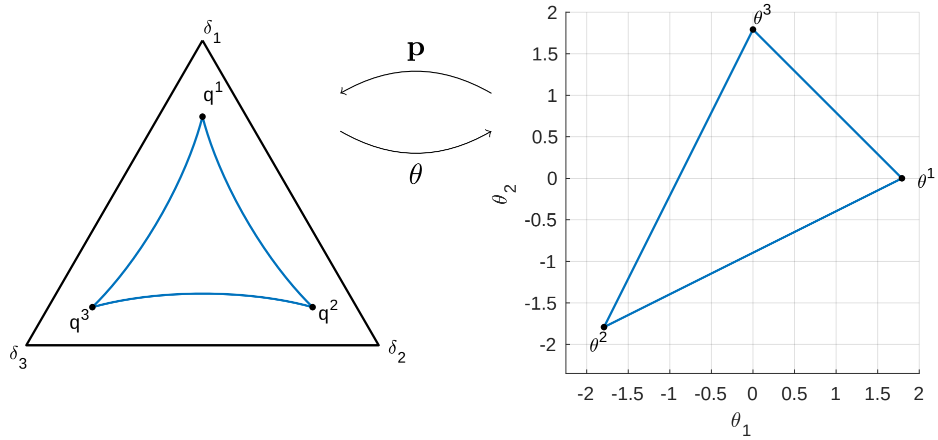

FIGURE 2. Exponential geodesic triangle plotted in the probability simplex (left) and in exponential parameter space (right). Exponential geodesics correspond to straight lines in exponential parameter space.

with vertices I, edges $E = \{\{a,b\},\{b,d\},\{a,c\},\{c,d\}\}$, and weights $\omega = (\omega_{ab}, \omega_{bd}, \omega_{ac}, \omega_{cd}) \in \mathbb{R}^E$ attached to the edges. The edge weights correspond to the inverse squared ground metric that we assign to the sample space I. The probability simplex for this sample space is the tetrahedron

$$\mathcal{P}(I) = \left\{ (p(x))_{x\in I} \in \mathbb{R}^4 \ : \ \sum_{x\in I} p(x) = 1, \quad p(x) \geq 0 \right\} .$$

Following Definition 4, the Wasserstein metric tensor is given by $g_p^W = L(p)^\dagger$, which is the inverse of the linear weighted Laplacian metric L from Definition 3. In this example the latter is

$$L(p) = \begin{pmatrix} \omega_{ab}\frac{p_a+p_b}{2} + \omega_{ac}\frac{p_a+p_c}{2} & -\omega_{ab}\frac{p_a+p_b}{2} & -\omega_{ac}\frac{p_a+p_c}{2} & 0 \\ -\omega_{ab}\frac{p_a+p_b}{2} & \omega_{ab}\frac{p_a+p_b}{2} + \omega_{bd}\frac{p_b+p_d}{2} & 0 & -\omega_{bd}\frac{p_b+p_d}{2} \\ -\omega_{ac}\frac{p_a+p_c}{2} & 0 & \omega_{ac}\frac{p_a+p_c}{2} + \omega_{cd}\frac{p_c+p_d}{2} & -\omega_{cd}\frac{p_c+p_d}{2} \\ 0 & \omega_{bd}\frac{p_b+p_d}{2} & -\omega_{cd}\frac{p_c+p_d}{2} & \omega_{bd}\frac{p_b+p_d}{2} + \omega_{cd}\frac{p_c+p_d}{2} \end{pmatrix}.$$

The independence model consist of the joint distributions that satisfy $p(x_1,x_2) = p(x_1)p(x_2)$. This can be parametrized in terms of $\Theta = \{\xi = (\xi_1,\xi_2) \in [0,1]^2\}$, where $\xi_1 = p_1(x_1 = +1)$, $\xi_2 = p_2(x_2 = +1)$ describe the marginal probability distributions. The parametrization $\mathbf{p}\colon \Theta \to \mathcal{P}(I)$ is then

$$\mathbf{p}(\xi)(x_1,x_2) = \begin{cases} (1-\xi_1)(1-\xi_2) & \textit{if } (x_1,x_2) = (-1,-1) \\ (1-\xi_1)\xi_2 & \textit{if } (x_1,x_2) = (-1,+1) \\ \xi_1(1-\xi_2) & \textit{if } (x_1,x_2) = (+1,-1) \\ \xi_1\xi_2 & \textit{if } (x_1,x_2) = (+1,+1) \end{cases} .$$

The model $\mathbf{p}(\Theta) \subset \mathcal{P}(I)$ is a two dimensional manifold. The parameter space Θ inherits the Riemannian structure g^W from $\mathcal{P}(I)$, which is computed as follows. Denote the Jacobi *matrix of the parametrization by*

$$J_\xi \mathbf{p}(\xi) = \begin{pmatrix} -(1-\xi_2) & -(1-\xi_1) \\ -\xi_2 & 1-\xi_1 \\ 1-\xi_2 & -\xi_1 \\ \xi_2 & \xi_1 \end{pmatrix} \in \mathbb{R}^{4\times 2} \ .$$

Then g^W *induces a metric tensor on* Θ *given by*

$$G(\xi) = J_\xi \mathbf{p}(\xi)^\mathsf{T} L(\mathbf{p}(\xi))^\dagger J_\xi(\mathbf{p}(\xi)) \in \mathbb{R}^{2\times 2} \ .$$

We now consider a discrete optimization problem via stochastic relaxation and illustrate the gradient flow. Consider following potential function on I*, taken from* [24]*:*

$$f(x_1,x_2) = x_1 + 2x_2 + 3x_1x_2 = \begin{cases} 0 & \textit{if } (x_1,x_2)=(-1,-1) \\ -2 & \textit{if } (x_1,x_2)=(-1,+1) \\ -4 & \textit{if } (x_1,x_2)=(+1,-1) \\ 6 & \textit{if } (x_1,x_2)=(+1,+1) \end{cases} .$$

We are to minimize $F(\mathbf{p}) = \mathbb{E}_\mathbf{p}[f]$*, i.e.,*

$$F(\mathbf{p}(\xi)) = \sum_{(x_1,x_2)\in I} f(x_1,x_2)p_1(x_1)p_2(x_2) = -4\xi_1 - 2\xi_2 + 12\xi_1\xi_2 \ .$$

By Theorem 7, the Wasserstein gradient flow is

$$\dot{\xi} = -G(\xi)^{-1}\nabla_\xi F(\mathbf{p}(\xi)) \ .$$

For our function, the standard Euclidean gradient is $\nabla_\xi F(\mathbf{p}(\xi)) = (-4+12\xi_2, -2+12\xi_1)^\mathsf{T}$*. The inverse matrix* G *is computed numerically from* J *and* L*.*

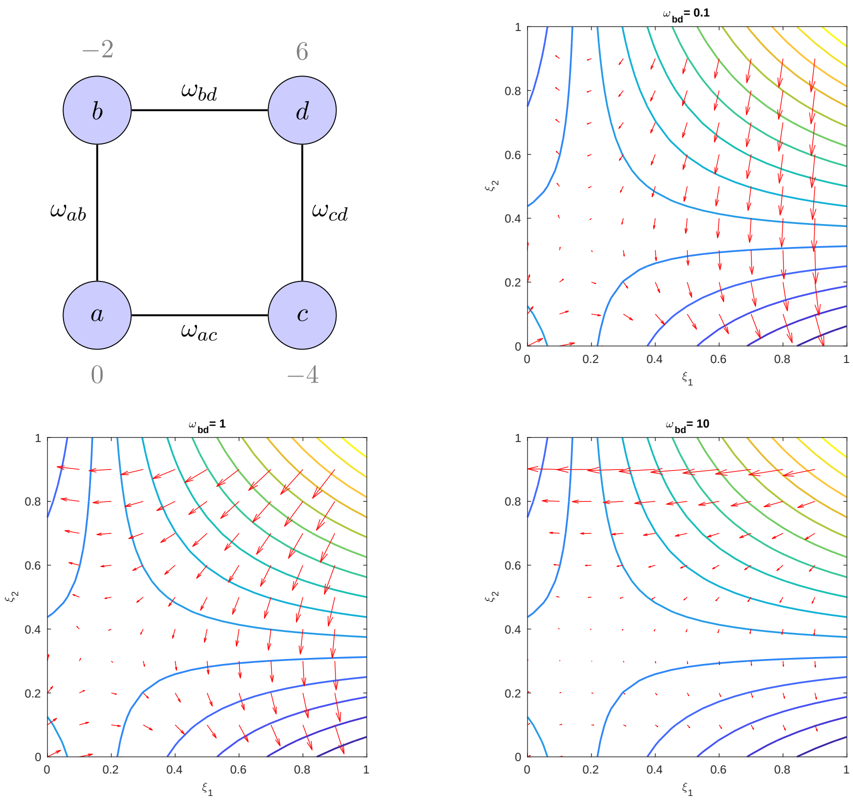

FIGURE 3. Negative Wasserstein gradient on the parameter space $[0,1]^2$ of the two-bit independence model from Example 2. We fix the state graph shown on the top left, and a function f with values shown in gray next to the state nodes. We evaluate the gradient flow for three different choices of the graph weight ω_{bd}. When the weight ω_{bd} is small, the flow from d towards b (a local minimum) is suppressed. A large weight has the opposite effect. The contours are for the objective function $F(\mathbf{p}(\xi)) = \mathbb{E}_{\mathbf{p}(\xi)}[f]$.

In Figure 3 we plot the negative Wasserstein gradient vector field in the parameter space $\Theta = [0,1]^2$. As can be seen, the Wasserstein gradient direction depends on the ground metric on sample space (encoded in the edge weights). If b and d are far away, there is higher tendency to to go c, rather than b. This reflects the intuition that, the more ground distance between b and d, the harder for the probability distribution to move from its concentration place b to d. We observe that the the attraction region of the two local minimizers changes dramatically as the ground metric between b and d changes, i.e., as ω_{bd} varies from 0.1, 1, 10.

This is different in the Fisher-Rao gradient flow, plotted in Figure 4, which is independent of the ground metric on sample space.

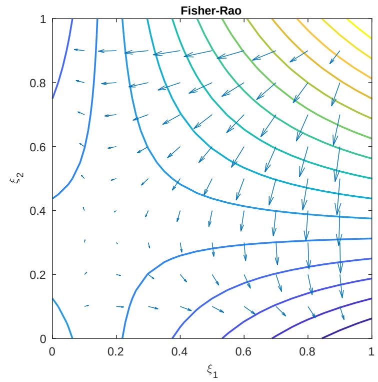

FIGURE 4. Fisher-Rao gradient vector field for the same objective function of Figure 3.

5. DISCUSSION

We introduced the Wasserstein statistical manifolds, which are submanifolds of the probability simplex with the L^2-Wasserstein Riemannian metric tensor. With this, we defined an *optimal transport* natural gradient flow on parameter space.

We hope that this paper contributes to strengthening the emerging interactions between information geometry and optimal transport, in particular in relation to machine learning problems. We suggest that many studies from information geometry will have a natural analog or extension in the Wasserstein statistical manifold. Some questions to consider include the following.

Is it possible to characterize the Wasserstein metric on probability manifolds through an invariance requirement of Chentsov type? For instance, the work [28] formulates extensions of Markov embeddings for polytopes and weighted point configurations. Is there a weighted graph structure for which the corresponding Wasserstein metric recovers the Fisher metric?

One of the key limitations of natural gradient methods in machine learning applications is the computational cost. This difficulty will likely be inherited in the Wasserstein gradient flow. However, the key innovation coming from the Wasserstein gradient is that it incorporates a ground metric in sample space. This could be fixed or trained. This could have a positive effect in terms of generalization performance, as it provides means to introduce preferences in the hypothesis space. The specific form of such a regularization still needs to be developed and investigated. In this regard, a natural question is how to define natural ground metric notions.

The Wasserstein distance has already been shown to be useful in machine learning, for instance in training restricted Boltzmann machines and generative adversarial networks. We think that the Wasserstein natural gradient can be a useful method for a variety of minimization problems. We will continue to work on the above problems.

Appendix

In this appendix we review the equivalence of static and dynamical formulations of the L^2-Wasserstein metric formally. For more details see [34].

Consider the duality of linear programming.

$$\begin{aligned}\frac{1}{2}W(\rho^0,\rho^1)^2 &= \inf_{\pi\geq 0}\Big\{\int_\Omega\int_\Omega \frac{1}{2}d_\Omega(x,y)^2\pi(x,y)dxdy\colon \int_\Omega \pi dy=\rho^0(x),\ \int_\Omega \pi dx=\rho^1(y)\Big\}\\ &= \sup_{\Phi^1,\Phi^0}\Big\{\int_\Omega \Phi^1(y)\rho^1(y)dy-\int_\Omega \Phi^0(x)\rho^0(x)dx\colon \Phi^1(y)-\Phi^1(x)\leq \frac{1}{2}d_\Omega(x,y)^2\Big\}\ .\end{aligned} \tag{17}$$

By standard considerations, the supremum in the last formula is attained when

$$\Phi^1(y)=\sup_{x\in\Omega}\ \Phi^0(x)+\frac{1}{2}d_\Omega(x,y)^2\ . \tag{18}$$

This means that Φ^1, Φ^0 are related to the viscosity solution of the Hamilton-Jacobi equation on Ω:

$$\frac{\partial\Phi(t,x)}{\partial t}+\frac{1}{2}g_x^\Omega(\nabla\Phi(t,x),\nabla\Phi(t,x))=0\ , \tag{19}$$

with $\Phi^0(x)=\Phi(0,x)$, $\Phi^1(x)=\Phi(1,x)$. Hence (17) becomes

$$\frac{1}{2}W(\rho^0,\rho^1)^2=\sup_{\Phi}\Big\{\int_\Omega \Phi^1(x)\rho^1(x)-\Phi^0(x)\rho^0(x)dx\colon \frac{\partial\Phi(t,x)}{\partial t}+\frac{1}{2}g_x^\Omega(\nabla\Phi(t,x),\nabla\Phi(t,x))=0\Big\}.$$

By the duality of above formulas, we can obtain variational problem (1). In other words, consider the dual variable of $\Phi_t=\Phi(t,x)$ by the density path $\rho_t=\rho(t,x)$, then

$$\begin{aligned}&\frac{1}{2}W(\rho^0,\rho^1)^2\\ =&\sup_{\Phi_t}\inf_{\rho_t}\int_\Omega \Phi^1\rho^1-\Phi^0\rho^0dx-\int_0^1\int_\Omega \rho_t\big[\partial_t\Phi_t+\frac{1}{2}g_x^\Omega(\nabla\Phi_t,\nabla\Phi_t)dx\big]dt\\ =&\sup_{\Phi_t}\inf_{\rho_t}\int_\Omega \Phi^1\rho^1-\Phi^0\rho^0dx-\int_0^1\int_\Omega \rho_t\partial_t\Phi_tdxdt-\int_0^1\int_\Omega\frac{1}{2}g_x^\Omega(\nabla\Phi_t,\nabla\Phi_t)\rho_tdxdt\\ =&\sup_{\Phi_t}\inf_{\rho_t}\int_0^1\int_\Omega \partial_t\rho_t\Phi_t-g_x^\Omega(\nabla\Phi_t,\nabla\Phi_t)\rho_tdxdt+\int_0^1\int_\Omega\frac{1}{2}g_x^\Omega(\nabla\Phi_t,\nabla\Phi_t)\rho_tdxdt\\ =&\inf_{\rho_t}\sup_{\Phi_t}\int_0^1\int_\Omega \Phi_t(\partial_t\rho_t+\mathrm{div}(\rho\nabla\Phi_t))dt+\int_0^1\int_\Omega\frac{1}{2}g_x^\Omega(\nabla\Phi_t,\nabla\Phi_t)\rho_tdxdt\\ =&\inf_{\rho_t}\ \Big\{\int_0^1\int_\Omega\frac{1}{2}g_x^\Omega(\nabla\Phi_t,\nabla\Phi_t)\rho_tdxdt\colon \partial_t\rho_t+\mathrm{div}(\rho\nabla\Phi_t)=0,\ \rho_0=\rho^0,\ \rho_1=\rho^1\Big\}\ .\end{aligned}$$

The third equality is derived by integration by parts w.r.t. t and the fourth equality is by switching infimum and supremum relations and integration by parts w.r.t. x.

In the above derivations, the relation of Hopf-Lax formula (18) and Hamilton-Jacobi equation (19) plays a key role for the equivalence of static and dynamic formulations of the Wasserstein metric. This is also a consequence of the fact that the sample space Ω is a length space, i.e.,

$$d_\Omega(x,y)^2 = \inf_{\gamma(t)} \Big\{ \int_0^1 g^\Omega_{\gamma(t)}(\dot\gamma,\dot\gamma)dt \colon \gamma(0)=x,\ \gamma(1)=y \Big\} .$$

However, in a discrete sample space I, there is no path $\gamma(t)\in I$ connecting two discrete points. Thus the relation between (18) and (19) does not hold on I. This indicates that in discrete sample spaces, the Wasserstein metric in Definition (1) can be different from the one defined by linear programming (5). See many related discussions in [11, 22].

Acknowledgement The authors would like to thank Prof. Luigi Malagò for his inspiring talk at UCLA in December 2017.

References

[1] S. Amari. Neural learning in structured parameter spaces - natural Riemannian gradient. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 127–133. MIT Press, 1997.
[2] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.
[3] S. Amari. *Information Geometry and Its Applications*. Number volume 194 in Applied mathematical sciences. Springer, Japan, 2016.
[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] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv:1701.07875 [cs, stat]*, 2017.
[6] N. Ay, J. Jost, H. Lê, and L. Schwachhöfer. *Information Geometry*. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics. Springer International Publishing, 2017.
[7] 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.
[8] L. Campbell. An extended Čencov characterization of the information metric. *Proceedings of the American Mathematical Society*, 98:135–141, 1986.
[9] E. A. Carlen and W. Gangbo. Constrained Steepest Descent in the 2-Wasserstein Metric. *Annals of Mathematics*, 157(3):807–846, 2003.
[10] N. N. Čencov. *Statistical decision rules and optimal inference*, volume 53 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, R.I., 1982. Translation from the Russian edited by Lev J. Leifman.
[11] S.-N. Chow, W. Huang, Y. Li, and H. Zhou. Fokker–Planck Equations for a Free Energy Functional or Markov Process on a Graph. *Archive for Rational Mechanics and Analysis*, 203(3):969–1008, 2012.
[12] S.-N. Chow, W. Li, and H. Zhou. A discrete Schrodinger equation via optimal transport on graphs. *arXiv:1705.07583 [math]*, 2017.
[13] S.-N. Chow, W. Li, and H. Zhou. Entropy dissipation of Fokker-Planck equations on graphs. *arXiv:1701.04841 [math]*, 2017.
[14] F. R. K. Chung. *Spectral Graph Theory*. Number no. 92 in Regional conference series in mathematics. Published for the Conference Board of the mathematical sciences by the American Mathematical Society, Providence, R.I, 1997.
[15] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. Poggio. Learning with a Wasserstein Loss. *arXiv:1506.05439 [cs, stat]*, 2015.
[16] W. Gangbo, W. Li, and C. Mou. Geodesic of minimal length in the set of probability measures on graphs. *arXiv:1712.09266 [math]*, 2017.
[17] J. D. Lafferty. The density manifold and configuration space quantization. *Transactions of the American Mathematical Society*, 305(2):699–741, 1988.

[18] G. Lebanon. Axiomatic geometry of conditional models. *IEEE Transactions on Information Theory*, 51(4):1283–1294, 2005.

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

[20] W. Li, P. Yin, and S. Osher. Computations of Optimal Transport Distance with Fisher Information Regularization. *Journal of Scientific Computing*, 2017.

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

[22] J. Maas. Gradient flows of the entropy for finite Markov chains. *Journal of Functional Analysis*, 261(8):2250–2292, 2011.

[23] L. Malagò, M. Matteucci, and G. Pistone. Towards the geometry of estimation of distribution algorithms based on the exponential family. In *Proceedings of the 11th Workshop Proceedings on Foundations of Genetic Algorithms*, FOGA '11, pages 230–242, New York, NY, USA, 2011. ACM.

[24] L. Malagò and G. Pistone. Natural Gradient Flow in the Mixture Geometry of a Discrete Exponential Family. *Entropy*, 17(12):4215–4254, 2015.

[25] A. Mielke. A gradient structure for reaction–diffusion systems and for energy-drift-diffusion systems. *Nonlinearity*, 24(4):1329–1346, 2011.

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

[27] G. Montavon, K.-R. Müller, and M. Cuturi. Wasserstein Training of Restricted Boltzmann Machines. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3718–3726. Curran Associates, Inc., 2016.

[28] G. Montúfar, J. Rauh, and N. Ay. On the Fisher metric of conditional probability polytopes. *Entropy*, 16(6):3207–3233, 2014.

[29] E. Nelson. *Quantum Fluctuations*. Princeton series in physics. Princeton University Press, Princeton, N.J, 1985.

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

[31] R. Pascanu and Y. Bengio. Revisiting natural gradient for deep networks. In *International Conference on Learning Representations 2014 (Conference Track)*, Apr. 2014.

[32] J. Peters, S. Vijayakumar, and S. Schaal. Natural actor-critic. In J. Gama, R. Camacho, P. B. Brazdil, A. M. Jorge, and L. Torgo, editors, *Machine Learning: ECML 2005*, pages 280–291, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

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

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

[35] T.-K. L. Wong. Logarithmic divergences from optimal transport and R\'enyi geometry. *arXiv:1712.03610 [cs, math, stat]*, 2017.

[36] S. Yi, D. Wierstra, T. Schaul, and J. Schmidhuber. Stochastic search using the natural gradient. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1161–1168, New York, NY, USA, 2009. ACM.