Information Geometry (2018) 1:181–214
https://doi.org/10.1007/s41884-018-0015-3

RESEARCH PAPER

Natural gradient via optimal transport

Wuchen Li[1] · **Guido Montúfar**[2,3]

Received: 15 March 2018 / Revised: 27 August 2018 / Published online: 19 November 2018

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 illustrate the analysis of elementary exponential family examples and demonstrate an application of the Wasserstein natural gradient to maximum likelihood estimation.

Keywords Optimal transport · Information geometry · Wasserstein statistical manifold · Displacement convexity · Machine learning

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 [11] characterizes this Riemannian metric as the only one, up to scaling, that is invariant with respect to natural statistical embeddings by

✉ Wuchen Li
wcli@math.ucla.edu

Guido Montúfar
montufar@math.ucla.edu

1 Department of Mathematics, University of California, Los Angeles, USA

2 Department of Mathematics and Department of Statistics, University of California, Los Angeles, USA

3 Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

Markov morphisms (see also [9,21,32]). 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., [1,27,35,36,40]).

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 [5,16,31]. One distinct feature of optimal transport is that it provides a distance among histograms that incorporates a ground metric on sample space. 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 [20]. The gradient descent method in the density manifold, called Wasserstein gradient flow, has been widely studied in the literature; see [34,38] 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 sample space deriving from physics [33], fluid mechanics [10] and differential geometry [25]. 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 [12,26,29]. Following the settings from [13,14,17,22], 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.

We note that 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 [38]. 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 [8]. 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

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

true in the setting of discrete sample spaces. We note that for the static formulation, there is no Riemannian metric tensor for the discrete probability simplex. See [14,26] for a detailed discussion.

In the literature, the exploration of connections between optimal transport and information geometry was initiated in [4,18,39]. These works focus on the distance function induced by linear programming on discrete sample spaces. As we pointed out above, this approach can not cover the Riemannian and differential structures induced by optimal transport. In this paper, we use the dynamical formulation of optimal transport to define a Riemannian metric structure for general statistical manifolds. With this, we obtain a natural gradient operator, which can be applied to any optimization problem over a parameterized statistical model. In particular, it is applicable to maximum likelihood estimation. Other works have studied the Gaussian family of distributions with L^2-Wasserstein metric [30,37]. In that particular case, the constrained optimal transport metric tensor can be written explicitly and the corresponding density submanifold is a totally geodesic submanifold. In contrast to those works, our discussion is applicable to arbitrary parametric models.

This paper is organized as follows. In Sect. 2 we briefly review the Riemannian manifold structure in probability space introduced by optimal transport in the cases of continuous and discrete sample spaces. In Sect. 3 we introduce Wasserstein statistical manifolds by isometric embedding into the probability manifold, and in Sect. 4 we derive the corresponding gradient flows. In Sect. 5 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. We introduce the corresponding Riemannian structure for simplices of probability distributions with discrete support.

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 discrete state case. For related studies, we refer the reader to [20,38] 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 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 $\Pi(\rho^0, \rho^1)$ 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 dynamic 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 dynamic (2) formulations is well known (for continuous Ω). 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

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

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

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

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

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

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 a tangent vector V_Φ of the space of densities (for more details see [20,25]). 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 $g^W : T_\rho\mathcal{P}_+(\Omega) \times T_\rho\mathcal{P}_+(\Omega) \to \mathbb{R}$ 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 $\Phi(x), \tilde{\Phi}(x) \in \mathcal{F}(\Omega)/\mathbb{R}$. This 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 in $(\mathcal{P}_+(\Omega), g^W)$ in the sense of [8,25]. In literature [20], $(\mathcal{P}_+(\Omega), g^W)$ 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 a discrete space $I = \{1, \dots, n\}$.

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

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

where d^G_{ij} represents the distance or *ground metric* 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

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

Here $p = (p_1, \dots, p_n)$ is a probability vector with coordinates p_i corresponding to the probabilities assigned to each node $i \in I$. We denote the relative interior of

the probability simplex by $\mathcal{P}_+(I)$. This consists of the strictly positive probability distributions, $p \in \mathcal{P}(I)$ 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 a *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 $1/d^G$ for simplicity of notations. In this way, 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 vector fields $v_{ij}, \tilde{v}_{ij}$ at each state $i \in I$ by

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

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 a probability distribution p:

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

This 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, $\mathrm{div}_G : \mathbb{R}^{n\times n} \to \mathbb{R}^n$ maps any given vector field m on the graph G to a potential function

$$\mathrm{div}_G(m) = \left(\sum_{j\in N(i)} \sqrt{\omega_{ij}} m_{ji} \right)_{i\in I}.$$

Second, the probability weighted gradient vector field $m = p\nabla_G \Phi$ 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 $(i, j) \in E$.

We are now 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 $W : \mathcal{P}_+(I) \times \mathcal{P}_+(I) \to \mathbb{R}$ by

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

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

$$\dot{p}(t) + \mathrm{div}_G(p(t)\nabla_G \Phi(t)) = 0, \ p(0) = p^0, \ p(1) = p^1.$$

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 as 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

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

The reason for this in-equivalence is that the 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 geometry and discrete probability simplex

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 [22]. 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 [17].

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

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

where $[\Phi] = \{(\Phi_1 + c, \ldots, \Phi_n + c) \colon c \in \mathbb{R}\}$ 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 [12] it is shown that $V_\Phi \colon \mathcal{F}(I)/\mathbb{R} \to T_p\mathcal{P}_+(I)$ is a well defined map which is 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 [25].

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 V_Φ and $V_{\tilde\Phi} \in T_p\mathcal{P}_+(I)$ to

$$g_p^W(V_\Phi, V_{\tilde\Phi}) = (\nabla_G\Phi, \nabla_G\tilde\Phi)_p. \tag{6}$$

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 $I = \{1, \ldots, n\}$ and a weighted graph $G = (I, E, \omega)$, 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 called oriented incidence matrix [15], 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 [15], 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 $0 < \lambda_1(p) \le \cdots \le \lambda_{n-1}(p)$ 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

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

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 $\sigma = L(p)\Phi$, $\tilde{\sigma} = L(p)\tilde{\Phi}$, we see that

$$\sigma^{\mathsf{T}} L(p)^\dagger \tilde{\sigma} = \Phi^{\mathsf{T}} L(p) L(p)^\dagger L(p)\tilde{\Phi} = \Phi^{\mathsf{T}} L(p)\tilde{\Phi} = (\nabla_G \Phi, \nabla_G \tilde{\Phi})_p.$$

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

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

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

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

This defines a Wasserstein Riemannian structure on the probability simplex. For more details of Riemannian formulas see [22]. Following [20] 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 statistical manifold*, or Wasserstein manifold for short.

3 Wasserstein statistical 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 a statistical model.

3.1 Wasserstein statistical manifold

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

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

In the sequel we will assume that $\operatorname{rank}(J_\theta p(\theta)) = d$, 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 $p\colon (\Theta, g) \to (\mathcal{P}_+(I), g^W)$ is an isometric embedding:

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

Here $dp(\theta)(a) = \big(\sum_{j=1}^n \frac{\partial p_i(\theta)}{\partial \theta_j} a_j\big)_{i=1}^n = J_\theta p(\theta) a$, where $J_\theta p(\theta)$ is the Jacobi matrix of $p(\theta)$ with respect to θ. We arrive at the following definition.

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

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

where $J_\theta p(\theta) = (\frac{\partial p_i(\theta)}{\partial \theta_j})_{1\leq i\leq n, 1\leq j\leq d} \in \mathbb{R}^{n\times d}$ is the Jacobi matrix of the parametrization p, and $L(p(\theta))^\dagger \in \mathbb{R}^{n\times n}$ 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 $p(\Theta)$. For this reason, we call $p(\Theta)$, or (Θ, I, 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 $L(p)^\dagger$ 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 $\theta \in \Theta$, we have*

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

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

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

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

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

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

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

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

Left multiply u_0 into (7), we obtain

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

Thus $c = 0$, and (7) forms

$$J_\theta p(\theta) a = 0.$$

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

We illustrate some geometric calculations on parameter space (Θ, g). For simplicity of illustration, we assume $\Theta \subset \mathbb{R}^d$, and denote a matrix function $G(\theta) \in \mathbb{R}^{d\times d}$ with $g_\theta(\dot\theta, \dot\theta) = \dot\theta^\mathsf{T} G(\theta)\dot\theta$, i.e.,

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

Under this notation, given $\theta_0, \theta_1 \in \Theta$, the Riemannian distance on (Θ, g) is defined by the geometric action functional:

$$\mathrm{Dist}(\theta_0,\theta_1)^2 = \inf_{\theta(\cdot)\in C^1([0,1];\Theta)} \left\{ \int_0^1 \dot\theta(t)^\mathsf{T} G(\theta(t))\dot\theta(t) dt \ : \ \theta(0) = \theta_0, \ \theta(1) = \theta_1 \right\}. \tag{9}$$

Denote $\theta(t) = \theta_t$, and S_t is the Legendre transformation of $\dot{\theta}_t$ in (Θ, g), then the cotangent geodesic flow satisfies

$$\begin{cases} \dot{\theta}_t - G(\theta_t)^{-1} S_t = 0 \\ \dot{S}_t + \frac{1}{2} \frac{\partial}{\partial \theta} S_t^{\mathsf{T}} G(\theta_t)^{-1} S_t = 0. \end{cases} \tag{10}$$

It is worth recalling the following facts. If p is an identity map, then (10) 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 (10) the *continuity equation* and the *Hamilton-Jacobi equation on parameter space*.

3.2 Geometry calculations in statistical manifold

We next present the geometric formulas in a probability model. This approach connects the geometry formulas in the full probability set to the ones in a submanifold $(p(\Theta), g)$, and in the parameter space (Θ, g).

We first study the orthogonal projection operator from $(\mathcal{P}_+(I), g^W)$ to $(p(\Theta), g)$.

Theorem 7 *Given $\theta \in \Theta$, for any tangent vector $\sigma \in T_{p(\theta)}\mathcal{P}_+(I)$, there exists a unique orthogonal decomposition*

$$\sigma = \sigma^{\parallel} + \sigma^{\perp}, \tag{11}$$

with $\sigma^{\parallel} \in T_{p(\theta)} p(\Theta)$ and $\sigma^{\perp} \in N_{p(\theta)} p(\Theta)$, i.e., $g^W_{p(\theta)}(\sigma^{\parallel}, \sigma^{\perp}) = 0$. At each point $p(\theta)$, the projection matrix

$$H(p(\theta)) = J_\theta p(\theta) \big(J_\theta p(\theta)^{\mathsf{T}} L(p(\theta))^{\dagger} J_\theta p(\theta) \big)^{\dagger} J_\theta p(\theta)^{\mathsf{T}} L(p(\theta))^{\dagger} \in \mathbb{R}^{n \times n},$$

gives the decomposition by

$$\sigma^{\parallel} = H(p(\theta))\sigma, \quad \sigma^{\perp} = (\mathbb{I} - H(p(\theta)))\sigma,$$

where $\mathbb{I}$ is an identity matrix in $\mathbb{R}^{n \times n}$.

Proof We first prove that (11) is a decomposition. It is to check that $g^W_{p(\theta)}(\sigma^{\parallel}, \sigma^{\perp}) = 0$, i.e.

$$\begin{aligned}&\sigma^{\mathsf{T}} H(p(\theta))^{\mathsf{T}} L(p(\theta))^{\dagger}(\mathbb{I} - H(p(\theta)))\sigma\\ &= \sigma^{\mathsf{T}}\Big(H(p(\theta))^{\mathsf{T}} L(p(\theta))^{\dagger} H(p(\theta)) - H(p(\theta))^{\mathsf{T}} L(p(\theta))^{\dagger}\Big)\sigma = 0.\end{aligned}$$

Recall $G(\theta) = J_\theta p(\theta)^{\mathsf{T}} L(p(\theta))^{\dagger} J_\theta p(\theta)$. We check that

$$\begin{aligned}&H(p(\theta))^{\mathsf{T}} L(p(\theta))^{\dagger} H(p(\theta))\\ &= L(p(\theta))^{\dagger} J_\theta p(\theta) G(\theta)^{\dagger} J_\theta p(\theta)^{\mathsf{T}} L(p(\theta))^{\dagger} J_\theta p(\theta) G(\theta)^{\dagger} J_\theta p(\theta)^{\mathsf{T}} L(p(\theta))^{\dagger}\\ &= L(p(\theta))^{\dagger} J_\theta p(\theta) G(\theta)^{\dagger} G(\theta) G(\theta)^{\dagger} J_\theta p(\theta)^{\mathsf{T}} L(p(\theta))^{\dagger}\\ &= L(p(\theta))^{\dagger} J_\theta p(\theta) G(\theta)^{\dagger} J_\theta p(\theta)^{\mathsf{T}} L(p(\theta))^{\dagger}\\ &= H(p(\theta))^{\mathsf{T}} L(p(\theta))^{\dagger},\end{aligned}$$

which shows the claim. We next prove the uniqueness of decomposition (11). Suppose there are two decomposition $\sigma = \sigma^{||} + \sigma^{\perp}$, $\tilde{\sigma} = \tilde{\sigma}^{||} + \tilde{\sigma}^{\perp}$, where $\sigma^{||} = J_\theta p(\theta)\dot{\theta}$ and $\tilde{\sigma}^{||} = J_\theta p(\theta)\dot{\tilde{\theta}}$. From the definition, then

$$\begin{aligned}0 &= g^W_p(\sigma^{||} - \tilde{\sigma}^{||}, \tilde{\sigma}^{\perp} - \sigma^{\perp}) = g^W_p(\sigma^{||} - \tilde{\sigma}^{||}, \sigma^{||} - \tilde{\sigma}^{||})\\ &= (J_\theta p(\theta)\dot{\theta} - J_\theta p(\theta)\dot{\tilde{\theta}})^{\mathsf{T}} L(p(\theta))^{\dagger}(J_\theta p(\theta)\dot{\theta} - J_\theta p(\theta)\dot{\tilde{\theta}})\\ &= (\dot{\theta} - \dot{\tilde{\theta}})^{\mathsf{T}} J_\theta p(\theta)^{\mathsf{T}} L(p(\theta))^{\dagger} J_\theta p(\theta)(\dot{\theta} - \dot{\tilde{\theta}})\\ &= (\dot{\theta} - \dot{\tilde{\theta}})^{\mathsf{T}} G(\theta)(\dot{\theta} - \dot{\tilde{\theta}}).\end{aligned}$$

Since $G(\theta)$ is positive definite, we have $\dot{\theta} = \dot{\tilde{\theta}}$ and $\sigma^{||} = \tilde{\sigma}^{||}$, which finishes the proof. □

We next present the second fundamental form for submanifold $(p(\Theta), g)$. Given any $\sigma, \tilde{\sigma} \in T_{p(\theta)} p(\Theta)$, consider the orthogonal decomposition of Levi–Civita connection in $(\mathcal{P}_+(I), g^W)$:

$$\nabla^W_\sigma \tilde{\sigma} = (\nabla^W_\sigma \tilde{\sigma})^{\parallel} + (\nabla^W_\sigma \tilde{\sigma})^{\perp}.$$

The second fundamental form is the orthogonal part of this decomposition, i.e., $B_{p(\theta)}(\sigma, \tilde{\sigma}) := (\nabla^W_\sigma \tilde{\sigma})^{\perp}$.

Proposition 8 *(Second fundamental form) Let* $\nabla_G \cdot \circ \nabla_G \cdot : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ *so that, for any* $\Phi, \tilde{\Phi} \in \mathbb{R}^n$,

$$(\nabla_G \Phi \circ \nabla_G \tilde{\Phi}) := \Big(g^I_i(\nabla_G \Phi, \nabla_G \tilde{\Phi})\Big)^n_{i=1} = \frac{1}{2}\Big(\sum_{j \in N(i)} \omega_{ij}(\Phi_i - \Phi_j)(\tilde{\Phi}_i - \tilde{\Phi}_j)\Big)^n_{i=1}.$$

Then

$$B_{p(\theta)}(\sigma,\tilde{\sigma}) = -\frac{1}{2}\Big(\mathbb{I} - H(p(\theta))\Big)\Big\{L(\sigma)L(p(\theta))^{\dagger}\tilde{\sigma} + L(\tilde{\sigma})L(p(\theta))^{\dagger}\sigma - L(p(\theta))(\nabla_G L(p(\theta))^{\dagger}\sigma \circ \nabla_G L(p(\theta))^{\dagger}\tilde{\sigma})\Big\}.$$

Proof As shown in [22, Proposition 11], the Christoffel formula in $(\mathcal{P}_+(I), g^W)$ satisfies

$$\nabla^W_{\sigma}\tilde{\sigma} = \frac{1}{2}\Big\{L(\sigma)L(p(\theta))^{\dagger}\tilde{\sigma} + L(\tilde{\sigma})L(p(\theta))^{\dagger}\sigma - L(p(\theta))(\nabla_G L(p(\theta))^{\dagger}\sigma \circ \nabla_G L(p(\theta))^{\dagger}\tilde{\sigma})\Big\}. \tag{12}$$

Following the projection operator $H(p(\theta))$, we finish the proof. □

We next establish the parallel transport and geodesic equation in $(p(\Theta), g)$.

Proposition 9 (Parallel transport) *Let $p(\theta_t) \in p(\Theta)$, $t \in (0,1)$ be a smooth curve. Consider a vector field $\sigma_t \in T_{p(\theta_t)}p(\Theta)$ along curve $p(\theta_t)$. Then the equation for σ_t to be parallel along $p(\theta_t)$ satisfies*

$$\dot{\sigma}_t = \frac{1}{2}H(p(\theta_t))\Big\{L(\sigma)L(p(\theta_t))^{\dagger}\dot{p}(\theta_t) + L(\dot{p}(\theta_t))L(p(\theta_t))^{\dagger}\sigma_t - L(p(\theta_t))(\nabla_G L(p(\theta_t))^{\dagger}p(\theta_t) \circ \nabla_G L(p(\theta_t))^{\dagger}\dot{p}(\theta_t))\Big\}.$$

If $\sigma_t = \dot{p}(\theta_t)$, then the geodesic equation satisfies

$$\ddot{p}(\theta_t) = H(p(\theta_t))\Big\{L(\dot{p}(\theta_t))L(p(\theta_t))^{\dagger}\dot{p}(\theta_t) - \frac{1}{2}L(p(\theta_t))(\nabla_G L(p(\theta_t))^{\dagger}p(\theta_t) \circ \nabla_G L(p(\theta_t))^{\dagger}\dot{p}(\theta_t))\Big\}.$$

Proof The parallel equation in a submanifold is given by

$$\dot{\sigma}_t + \Big(\nabla^W_{\dot{p}(\theta_t)}\sigma_t\Big)^{\parallel} = 0.$$

In other words, we have

$$\dot{\sigma}_t = -H(p(\theta_t))\nabla^W_{\dot{p}(\theta_t)}\sigma_t,$$

where ∇^W is defined in (12). Let $\sigma_t = \dot{p}(\theta_t)$, then

$$\ddot{p}(\theta_t) + \Big(\nabla^W_{\dot{p}(\theta_t)}\dot{p}(\theta_t)\Big)^{\parallel} = 0.$$

This means that

$$\ddot{p}(\theta_t) = -H(\theta_t)\nabla^W_{\dot{p}(\theta_t)}\dot{p}(\theta_t).$$

Following the projection operator and (12), we finish the proof. □

We last present the curvature tensor in $(p(\Theta), g)$, denoted by $R(\cdot,\cdot)\cdot\colon T_{p(\Theta)}p(\Theta)\times T_{p(\theta)}p(\Theta)\times T_{p(\theta)}p(\Theta)\to T_{p(\theta)}p(\Theta)$.

Proposition 10 (Curvature tensor) *Given σ_1, σ_2, σ_3, $\sigma_4\in T_{p(\theta)}p(\Theta)$, then*

$$\begin{aligned}
g_{p(\theta)}(R(\sigma_1,\sigma_2)\sigma_3,\sigma_4) &= m(\sigma_1,\sigma_4)^T\Big(\mathbb{I}-H(p(\theta))\Big)^T L(p(\theta))^\dagger\Big(\mathbb{I}-H(p(\theta))\Big)m(\sigma_2,\sigma_3)\\
&\quad - m(\sigma_1,\sigma_3)^T\Big(\mathbb{I}-H(p(\theta))\Big)^T L(p(\theta))^\dagger\Big(\mathbb{I}-H(p(\theta))\Big)m(\sigma_2,\sigma_4)\\
&\quad + \frac{1}{2}\Big\{\sigma_2^T L(p(\theta))^\dagger L(m(\sigma_1,\sigma_3))L(p(\theta))^\dagger\sigma_4\\
&\quad + \sigma_1^T L(p(\theta))^\dagger L(m(\sigma_2,\sigma_4))L(p(\theta))^\dagger\sigma_3\\
&\quad - \sigma_2^T L(p(\theta))^\dagger L(m(\sigma_1,\sigma_4))L(p(\theta))^\dagger\sigma_3\\
&\quad - \sigma_1^T L(p(\theta))^\dagger L(m(\sigma_2,\sigma_3))L(p(\theta))^\dagger\sigma_4\Big\}\\
&\quad + \frac{1}{4}\Big\{2n(\sigma_1,\sigma_2)^T L(p(\theta))^\dagger n(\sigma_3,\sigma_4)\\
&\quad + n(\sigma_1,\sigma_3)^T L(p(\theta))^\dagger n(\sigma_2,\sigma_4)\\
&\quad - n(\sigma_2,\sigma_3)^T L(p(\theta))^\dagger n(\sigma_1,\sigma_4)\Big\},
\end{aligned}$$

where m, $n\colon T_{p(\theta)}p(\Theta)\times T_{p(\theta)}p(\Theta)\to T_{p(\theta)}p(\Theta)$ are symmetric, antisymmetric operators respectively, which are defined by

$$\begin{aligned}
m(\sigma_a,\sigma_b) &:= \nabla^W_{\sigma_a}\sigma_b = \frac{1}{2}\Big\{L(\sigma_a)L(p(\theta))^\dagger\sigma_b + L(\sigma_b)L(p(\theta))^\dagger\sigma_a\\
&\quad - L(p(\theta))(\nabla_G L(p(\theta))^\dagger\sigma_a\circ\nabla_G L(p(\theta))^\dagger\sigma_b)\Big\},
\end{aligned}$$

and

$$n(\sigma_a,\sigma_b) := L(\sigma_a)L(p(\theta))^\dagger\sigma_b - L(\sigma_b)L(p(\theta))^\dagger\sigma_a.$$

Proof The curvature tensor in submanifold relates to the one in full manifold as follows:

$$\begin{aligned}
g_{p(\theta)}(R(\sigma_1,\sigma_2)\sigma_3,\sigma_4) &= B_{p(\theta)}(\sigma_1,\sigma_4)^{\mathsf{T}}L(p(\theta))^\dagger B_{p(\theta)}(\sigma_2,\sigma_3)\\
&\quad - B_{p(\theta)}(\sigma_1,\sigma_3)^{\mathsf{T}}L(p(\theta))^\dagger B_{p(\theta)}(\sigma_2,\sigma_4)\\
&\quad + g_{p(\theta)}(R_W(\sigma_1,\sigma_2)\sigma_3,\sigma_4),
\end{aligned}$$

where R_W is the curvature tensor of $(\mathcal{P}_+(I), g^W)$ derived in [22, Proposition 6]. Combining R_W and the second fundamental form in Proposition 8, we derive the result. □

4 Gradient flow on Wasserstein statistical manifold

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

4.1 Gradient flow on parameter space

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

$$g_\theta(\nabla_g F(p(\theta)), a) = \nabla_\theta F(p(\theta)) \cdot a, \quad \text{for any } a \in T_\theta\Theta, \tag{13}$$

where $\nabla_\theta F(p(\theta)) \cdot a = \sum_{i=1}^d \frac{\partial}{\partial\theta_i} F(p(\theta))a_i$. The gradient flow satisfies

$$\dot\theta_t = -\nabla_g F(p(\theta_t)).$$

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

Theorem 11 (Wasserstein gradient flow) *The gradient flow of a functional* $F\colon \mathcal{P}_+(I) \to \mathbb{R}$ *is given by*

$$\dot\theta_t = -G(\theta_t)^{-1}\nabla_\theta F(p(\theta_t)),$$

where ∇_θ *is the Euclidean gradient of* $F(p(\theta))$ *with respect to* θ. *More explicitly,*

$$\dot\theta_t = -\Big(J_\theta p(\theta_t)^T L(p(\theta_t))^\dagger J_\theta p(\theta_t)\Big)^\dagger J_\theta p(\theta_t)^T \nabla_p F(p(\theta_t)), \tag{14}$$

where ∇_p *is the Euclidean gradient of* $F(p)$ *with respect to* p.

Proof The proof follows directly from (13). Notice that

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

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

$$\nabla_g F(p(\theta)) = \Big(J_\theta p(\theta)^\mathsf{T} L(p(\theta))^\dagger J_\theta p(\theta)\Big)^\dagger \nabla_\theta F(p(\theta)).$$

We compute $\nabla_\theta F(p(\theta))$ as

$$\nabla_\theta F(p(\theta)) = \left(\frac{\partial}{\partial\theta_i} F(p(\theta))\right)_{i=1}^n = \left(\sum_{j=1}^n \frac{\partial}{\partial p_j} F(p(\theta)) \cdot \frac{\partial p_j(\theta)}{\partial\theta_i}\right)_{i=1}^n = J_\theta p(\theta)^\mathsf{T} \nabla_p F(p(\theta)).$$

This concludes the proof of (14). □

Equation (14) is the generalization of Wasserstein gradient flow in probability simplex to the one on parameter space. If p is an identity map with the parameter space Θ equal to the entire probability simplex, then (14) is

$$\dot p_t = -\nabla_g F(p_t) = \mathrm{div}_G(p_t \nabla_G \nabla_p F(p_t)),$$

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

$$\partial_t \rho_t = -\nabla_W F(\rho_t) = \mathrm{div}(\rho_t \nabla \delta_\rho F(\rho_t)),$$

which is the Wasserstein gradient flow on Ω. From now on, we call (14) the *Wasserstein gradient flow on parameter space*.

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

$$\arg\min_{h \in T_\theta \Theta} F(p(\theta + h)) \quad \text{s.t.} \quad \frac{1}{2} W(p(\theta), p(\theta + h))^2 = \epsilon, \tag{15}$$

where $\epsilon \in \mathbb{R}_+$ is a given small constant. By taking the second-order Taylor approximation of the Wasserstein distance at θ, we get

$$W(p(\theta), p(\theta + h))^2 = h^\mathsf{T} G(\theta) h + o(h^2),$$

where $G(\theta)$ is the metric tensor of (Θ, g) defined in (8), inherited from Wasserstein manifold. We take the first-order approximation of $F(p(\theta + h))$ in (15) by

$$\arg\min_{h \in T_\theta \Theta} F(p(\theta)) + h^\mathsf{T} \nabla_\theta F(p(\theta)) \quad \text{s.t.} \quad \frac{1}{2} h^\mathsf{T} G(\theta) h = \epsilon.$$

By the Lagrangian method with Lagrange multiplier λ, we have

$$h = \lambda G(\theta)^{-1} \nabla_\theta F(p(\theta)).$$

The above derivations lead to the Wasserstein natural gradient direction

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

Remark 2 In the standard Fisher–Rao natural gradient [2], we replace (15) by

$$\arg\min_h F(p(\theta + h)) \quad \text{s.t.} \quad \mathrm{KL}(p(\theta) \| p(\theta + h)) = \epsilon,$$

where KL stands for the Kullback-Leibler divergence (relative entropy) from $p(\theta)$ to $p(\theta + h)$. Our definition changes the KL-divergence by the Wasserstein distance.

4.2 Displacement convexity on parameter space

The Wasserstein structure on the statistical manifold not only provides us the gradient operator, but also the Hessian operator on (Θ, g). The latter allows us to introduce the displacement convexity on parameter space.

We first review some facts. One remarkable property of Wasserstein geometry is that it yields a correspondence between differential operators on sample space and

differential operators on probability space. E.g., the Hessian operator on Wasserstein manifold is equal to the expectation of Hessian operator on sample space.

An important example is stochastic relaxation. Given $f(x) \in C^\infty(\Omega)$, consider

$$F(\rho) = \mathbb{E}_{X\sim\rho}[f(X)] = \int_\Omega f(x)\rho(x)dx.$$

It is known that the Hessian operator of $F(\rho)$ on Wasserstein manifold satisfies

$$\mathrm{Hess}_W F(\rho)(V_\Phi, V_{\tilde\Phi}) = \mathbb{E}_{X\sim\rho}(\mathrm{Hess}\, f(X)\nabla\Phi(X), \nabla\tilde\Phi(X)).$$

One can show that $\mathrm{Hess}\, f \succeq \lambda\mathbb{I}$ if and only if $\mathrm{Hess}_W F(\rho)(V_\Phi, V_\Phi) \geq \lambda g^W_\rho(V_\Phi, V_\Phi)$. This means that f is λ-geodesic convex in (Ω, g^Ω) if and only if $F(\rho)$ is λ-geodesic convex in $(\mathcal{P}(\Omega), g^W)$. In literature [38], the geodesic convexity on Wasserstein manifold is known as the displacement convexity.

In this section we would like to extend the displacement convexity to parameter space Θ. In other words, we relate the parameter to the differential structures of sample space via constrained Wasserstein geometry (Θ, g). If Θ is the full probability manifold, our definition coincides with the classical Hessian operator in sample space.

Definition 12 (*Displacement convexity on parameter space*) Given $F \circ p\colon \Theta \to \mathbb{R}$, we say that $F(p(\theta))$ is λ-displacement convex if for any constant speed geodesic θ_t, $t \in [0, 1]$ connecting $\theta_0, \theta_1 \in (\Theta, g)$, it holds that

$$F(p(\theta_t)) \leq (1-t)F(p(\theta_0)) + tF(p(\theta_1)) - \frac{\lambda}{2}t(1-t)\,\mathrm{Dist}(\theta_0, \theta_1)^2,$$

where Dist is defined in (9). If $F(p(\theta)) = \sum_{i=1}^n f_i p_i(\theta)$ is λ-displacement convex, then we call $f \in \mathbb{R}^n$ λ-convex in (Θ, I, p).

Remark 3 In particular, the displacement convexity of KL divergence relates to the Ricci curvature lower bound on sample space. We elaborate this notion in [23].

We next derive the displacement convexity condition for stochastic relaxation.

Theorem 13 *Assume $\Theta \subset \mathbb{R}^d$ is a compact set and $f = (f_i)_{i=1}^n \in \mathbb{R}^n$. Then f is λ-convex if and only if*

$$\sum_{i=1}^n p_i(\theta)\Big(\Gamma(\Gamma(f, \Phi), \Phi) - \frac{1}{2}\Gamma(\Gamma(\Phi, \Phi), f)\Big)_i + \sum_{i=1}^n f_i B_{p(\theta)}(V_\Phi, V_\Phi)_i$$
$$\geq \lambda \sum_{i=1}^n \Gamma(\Phi, \Phi)_i p_i(\theta), \tag{16}$$

for any $\Phi \in \mathcal{F}(I)/\mathbb{R}$ and $\theta \in \Theta$. Here $\Gamma\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is given by

$$\Gamma(\Phi, \tilde\Phi)_i := g^I_i(\nabla_G\Phi, \nabla_G\tilde\Phi) = \frac{1}{2}\sum_{j\in N(i)} \omega_{ij}(\Phi_i - \Phi_j)(\tilde\Phi_i - \tilde\Phi_j),$$

and B is the second fundamental form given in Proposition 8.

Proof If Θ is a compact set, then the λ-displacement convexity of $F(p(\theta))$ is equivalent to

$$\mathrm{Hess}_g\, F(p(\theta)) \succeq \lambda G(\theta),$$

where $\mathrm{Hess}_g\, F$ is the Hessian operator in (Θ, g). We next calculate this Hessian operator explicitly. Notice that

$$\mathrm{Hess}_g\, F(\sigma, \tilde{\sigma}) = \mathrm{Hess}_W\, F(\sigma, \tilde{\sigma}) + B_{p(\theta)}(\sigma, \tilde{\sigma})^{\mathsf{T}} \nabla_p F(p(\theta)),$$

where Hess_W is the Hessian operator in $(\mathcal{P}_+(I), g^W)$. Denote the above in dual coordinates, i.e. $\sigma = \tilde{\sigma} = V_\Phi = V_{\tilde{\Phi}} = L(p(\theta))\Phi$, and follow the geometric computations in [22, Proposition 18], we finish the proof. □

Here Γ is the discrete Bakry-Emery Gamma one operator [7]. The geometry of Wasserstein manifold is directly related to the expectation of Bakry-Emery Gamma one operators [22]. In particular, if p is the identity mapping and $I = \Omega$, then our definition (16) represents

$$\int_\Omega \Big(\Gamma(\Gamma(f, \Phi), \Phi) - \frac{1}{2}\Gamma(\Gamma(\Phi, \Phi), f)\Big)\rho(x)dx \geq \lambda \Gamma(\Phi, \Phi)\rho dx,$$

i.e.

$$\int_\Omega \mathrm{Hess}\, f(x)(\nabla\Phi(x), \nabla\Phi(x))\rho(x)dx \geq \lambda \int_\Omega g_x^\Omega(\nabla\Phi, \nabla\Phi)\rho(x)dx$$

for any ρ, and vector fields $\nabla\Phi$. The above inequality is same as requiring $\mathrm{Hess}\, f \succeq \lambda I$. Our definition extends this concept to parameter space.

4.3 Numerical methods

Here we discuss the numerical computation of the Wasserstein metric and the Wasserstein gradient flow.

Let us give a simple reformulation of the gradient that can be useful in practice, where typically $n \gg d$. Note that

$$\Big(J_\theta p(\theta)^{\mathsf{T}} L(p(\theta))^\dagger J_\theta p(\theta)\Big)^\dagger = J_\theta p(\theta)^\dagger L(p(\theta))(J_\theta p(\theta)^{\mathsf{T}})^\dagger.$$

Hence (7) can be written as

$$\frac{d\theta}{dt} = -J_\theta p(\theta)^\dagger L(p(\theta))(J_\theta p(\theta)^{\mathsf{T}})^\dagger J_\theta p(\theta)^{\mathsf{T}} \nabla_p F(p(\theta)\,.$$

In this formulation, the computation of the pseudo inverse of $L(p(\theta)) \in \mathbb{R}^{n\times n}$ is not needed, and the computation complexity reduces to that of obtaining the pseudo inverse of $J_\theta p(\theta) \in \mathbb{R}^{n\times d}$.

Natural Wasserstein gradient method

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

1. Choose a suitable step size $\lambda_k > 0$;
2. $\theta^{k+1} = \theta^k - \lambda_k \big((J_\theta p(\theta^k))^\mathsf{T} L(p(\theta^k))^\dagger J_\theta p(\theta^k)\big)^\dagger (J_\theta p(\theta^k))^\mathsf{T} \nabla_p F(p(\theta^k))$;

end

Natural Jordan–Kinderlehrer–Otto scheme

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

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

end

Given the gradient flow (7), there are two standard choices of time discretization, namely the forward Euler scheme and the backward Euler scheme. Denote the step size by $\lambda > 0$. The forward Euler method computes a discretized trajectory by

$$\theta^{k+1} = \theta^k - \lambda \nabla_g F(p(\theta^k)),$$

while the backward Euler method computes

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

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 is often called the Jordan–Kinderlehrer–Otto (JKO) scheme. In the following we give pseudo code for both numerical methods.

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. 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.

5 Examples

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 $\mathbb{R}^3$:

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

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

$$\begin{aligned} W(p^0, p^1)^2 \ := \inf_{\Phi(t)} \int_0^1 \Big\{ &(\Phi_1(t) - \Phi_2(t))^2 \frac{p_1(t) + p_2(t)}{2} \\ &+ (\Phi_2(t) - \Phi_3(t))^2 \frac{p_2(t) + p_3(t)}{2} \Big\} dt, \end{aligned} \tag{17}$$

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 (17). We parametrize a probability vector as $p = (p_1, 1 - p_1 - p_3, p_3)$, with parameters (p_1, p_3). Then (17) can be written as

$$W(p^0, p^1)^2 := \inf_{p(t)\colon\, p(0)=p^0,\ p(1)=p^1} \int_0^1 \Big\{ \frac{\dot{p}_1(t)^2}{1 - p_3(t)} + \frac{\dot{p}_3(t)^2}{1 - p_1(t)} \Big\} dt. \tag{18}$$

where the infimum is taken over paths $p\colon [0, 1] \to \mathcal{P}_+(I)$. We also compare the Wasserstein metric (18) with the Fisher–Rao metric. In this case, the Fisher–Rao metric function is given by

$$\mathrm{FR}(p^0, p^1)^2 := \inf_{p(t)\colon\, p(0)=p^0,\ p(1)=p^1} \int_0^1 \Big\{ \frac{\dot{p}_1(t)^2}{p_1(t)} + \frac{(\dot{p}_1(t) + \dot{p}_3(t))^2}{p_2(t)} + \frac{\dot{p}_3(t)^2}{p_3(t)} \Big\} dt.$$

This clearly demonstrates the difference between Wasserstein Riemannian metric and Fisher–Rao metric. We would also compare the dynamical optimal transport with the statistical one. In particular, if the ground metric is given by $c_{12} = 1, c_{13} = 2, c_{23} = 1$, which is of homogenous degree one type. Then the statistical optimal transport defined by

$$d(p^0, p^1) = \inf_{\pi \geq 0} \Big\{ c_{12}\pi_{12} + c_{13}\pi_{13} + c_{12}\pi_{23} \colon \sum_{i=1}^3 \pi_{ij} = p_j^0, \ \sum_{j=1}^3 \pi_{ij} = p_i^1 \Big\},$$

can be reformulated by

$$d(p^0, p^1) = \inf_{p(t)\colon\, p(0)=p^0,\ p(1)=p^1} \int_0^1 \Big\{ |\dot{p}_1(t)| + |\dot{p}_3(t)| \Big\} dt.$$

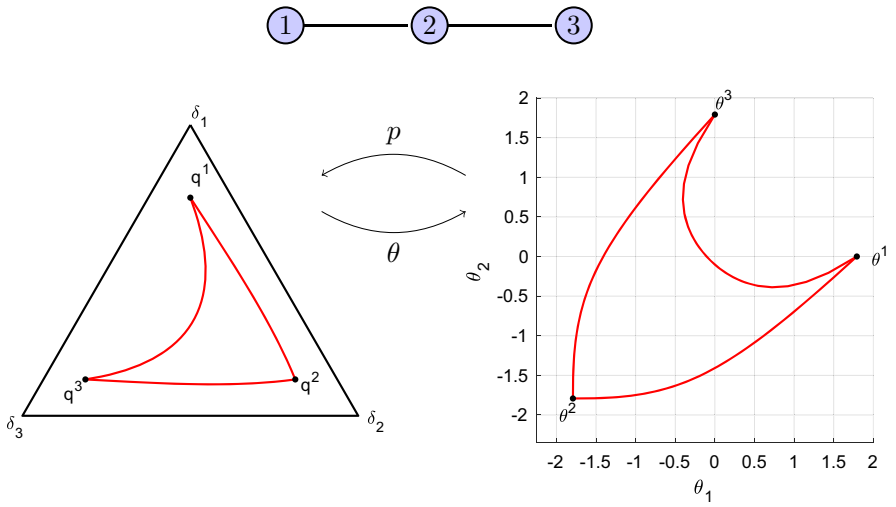

Fig. 1 The Wasserstein geodesic triangle from Example 1 plotted in the probability simplex (left) and in exponential parameter space (right). The path connecting q^1 and q^3 bends towards q^2; something that does not happen for the other two paths. This illustrates how, as a result of the ground metric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

Here the statistical formulation does not provide a Riemannian metric, but gives a Finslerian metric.

We next compute (18) numerically[2] for different choices of the boundary conditions p^0, p^1. We fix three distributions

$$q^1 = \frac{1}{8}(6,1,1), \quad q^2 = \frac{1}{8}(1,6,1), \quad q^3 = \frac{1}{8}(1,1,6) \tag{19}$$

and solve (18) for three choices of the boundary conditions:

$$p^0 = q^1,\ p^1 = q^2; \quad p^0 = q^1,\ p^1 = q^3; \quad p^0 = q^2,\ p^1 = q^3. \tag{20}$$

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

$$p_1(\theta) = \frac{e^{\theta_1}}{e^{\theta_1} + e^{\theta_2} + 1}, \quad p_3(\theta) = \frac{e^{\theta_2}}{e^{\theta_1} + e^{\theta_2} + 1},$$

[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 reference for these techniques is [24].

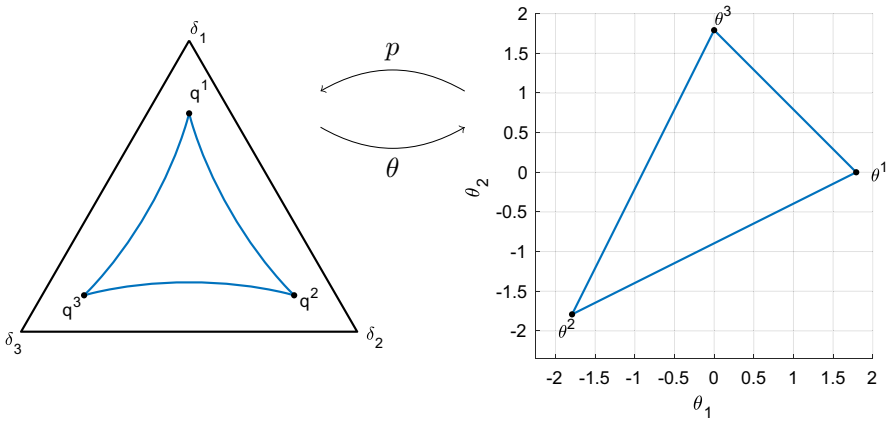

Fig. 2 Exponential geodesic triangle plotted in the probability simplex (left) and in exponential parameter space (right). Exponential geodesics correspond to straight lines in exponential parameter space

$$p_2(\theta) = 1 - p_1(\theta) - p_3(\theta) = \frac{1}{e^{\theta_1} + e^{\theta_2} + 1}.$$

We rewrite the Wasserstein metric (18) in terms of θ. Denote $p(\theta^k) = p^k$, $k = 0, 1$. Then the Wasserstein metric in the coordinate system θ is

$$\begin{aligned}&\mathrm{Dist}(\theta^0, \theta^1)^2\\ &= \inf_{\theta(t):\, \theta(0)=\theta^0,\ \theta(1)=\theta^1} \left\{ \int_0^1 \dot\theta^{\mathsf T} J_\theta(p_1, p_3)^{\mathsf T} \begin{pmatrix} \frac{1}{1-p_3(\theta)} & 0 \\ 0 & \frac{1}{1-p_1(\theta)} \end{pmatrix} J_\theta(p_1, p_3)\dot\theta dt \right\}.\end{aligned}$$

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

For comparison, we compute the *exponential* geodesic triangle between the same distributions q^1, q^2, q^3. This is shown in Fig. 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

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

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

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

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 $p\colon \Theta \to \mathcal{P}(I)$ is then

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

The model $p(\Theta) \subset \mathcal{P}(I)$ 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

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

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

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

We now consider a discrete optimization problem via stochastic relaxation and illustrate the gradient flow. Consider following potential function on I, taken from [28]:

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

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

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

By Theorem 11, the Wasserstein gradient flow is

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

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

In Fig. 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 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 Fig. 4, which is independent of the ground metric on sample space.

The above result illustrates the displacement convexity shown in Theorem 16. Different ground metric exhibits different displacement convexity of f on parameter space (Θ, g). These properties lead to different convergence regions of Wasserstein gradient flows.

Example 3 (Wasserstein gradient for maximum likelihood estimation) In maximum likelihood estimation, we seek to minimize the Kullback-Leibler divergence

$$\mathrm{KL}(q\,\|\,p(\theta)) = \sum_{x\in I} q_x \log\frac{q_x}{p_x(\theta)},$$

where q is the empirical distribution of some given data. The Wasserstein gradient flow of $\mathrm{KL}(q\,\|\,p(\theta))$ satisfies

$$\frac{d\theta}{dt} = \Big(J_\theta\, p(\theta)^{\mathsf{T}} L(p(\theta))^{\dagger} J_\theta\, p(\theta)\Big)^{\dagger} J_\theta\, p(\theta)^{\mathsf{T}} \left(\frac{q}{p(\theta)}\right).$$

In this example we consider hierarchical log-linear models as our parametrized probability models, which are an important type of exponential families describing interactions among groups of random variables. Concretely, for an inclusion closed set S of subsets of $\{1,\dots,n\}$, the hierarchical model $\mathcal{E}_S$ for n binary variables is the set of distributions of the form

$$p_x(\theta) = \frac{1}{Z(\theta)}\exp\Big(\sum_{\lambda\in S}\theta_\lambda\phi_\lambda(x)\Big), \quad x\in\{0,1\}^n,$$

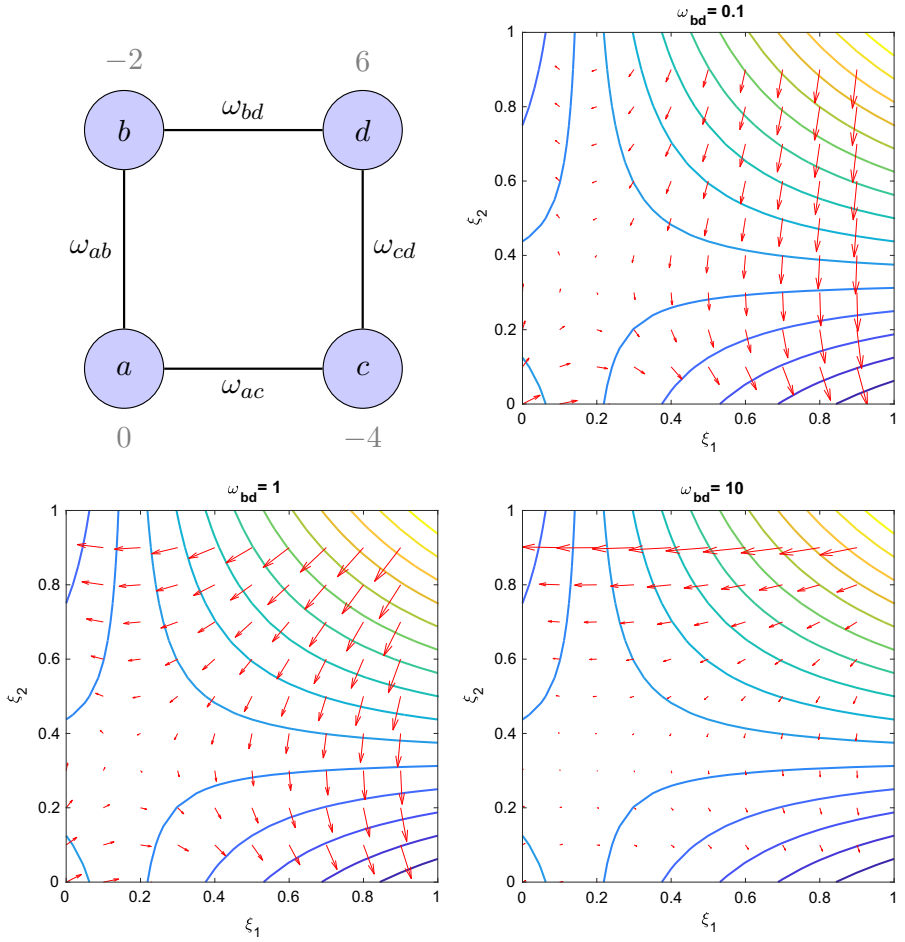

Fig. 3 Negative Wasserstein gradient on the parameter space $[0, 1]^2$ of the two-bit independence model from Example 2. We fix the state graph shown on the top left, and a function f with values shown in gray next to the state nodes. We evaluate the gradient flow for three different choices of the graph 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 $F(p(\xi)) = \mathbb{E}_{p(\xi)}[f]$

for all possible choices of parameters $\theta_\lambda \in \mathbb{R}$, $\lambda \in S$. Here the ϕ_λ are real valued functions with $\phi_\lambda(x) = \phi_\lambda(y)$ whenever $x_i = y_i$ for all $i \in \lambda$. We consider two different choices of ϕ_λ, $\lambda \in S$, corresponding to two different parametrizations of the model.

- Our first choice are the orthogonal characters

$$\sigma_\lambda(x) = \prod_{i\in\lambda}(-1)^{x_i} = e^{i\pi\langle 1_\lambda, x\rangle}, \quad x \in \{0, 1\}^n,$$

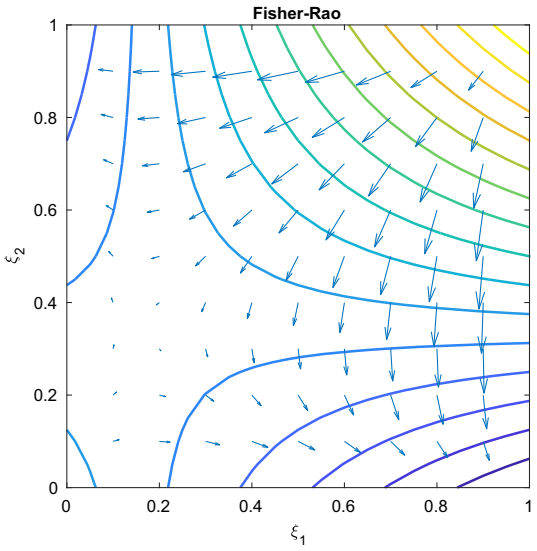

Fig. 4 Fisher–Rao gradient vector field for the same objective function of Fig. 3

which can be interpreted as a Fourier basis for the space of real valued functions over binary vectors.

- As an alternative choice we consider the basis of monomials

$$\pi_\lambda(x) = \prod_{i\in\lambda} x_i, \quad x \in \{0,1\}^n,$$

which is not orthogonal, but is frequently used in practice.

When $S = \{\lambda \subseteq \{1,\dots,n\} \colon |\lambda| \leq k\}$, the model is called k-interaction model. We consider k-interaction models with $k = 1,\dots,n$ (independence model, pair interaction model, three way interaction model, etc.), with the two parametrizations, σ (orthogonal sufficient statistics) and π (non-orthogonal sufficient statistics).

We compare the Euclidean, Fisher, and Wasserstein gradients. For binary variables, the Hamming distance is a natural ground metric notion. Accordingly, we define the Wasserstein metric with the uniformly weighted graph of the binary cube. We sampled a few target distributions on $\{0,1\}^n$ uniformly at random (uniform Dirichlet). For each target distribution, we initialize the model at the uniform distribution, $\theta_0 = 0$. The gradient descent parameter iteration is

$$\theta_{t+1} = \theta_t - \gamma_t G(\theta_t)^{-1} \nabla \mathrm{KL}(q \,\|\, p_{\theta_t}),$$

where G is the corresponding metric (Euclidean, Fisher, or Wasserstein), ∇ is the standard gradient operator with respect to the model parameter θ, and $\gamma_t \in \mathbb{R}_+$ is the learning rate (step size). The choice of the learning rate γ_t is important and the optimal value may vary for different methods and problems. We implemented an adaptive method to handle this as follows. We set an initial learning rate $\gamma_0 = 0.001$,

and at each iteration t, if the divergence does not decrease, we scale down the learning rate by a factor of 3/4. We also tried a few other methods, including backtracking line search and Adam [19], which is a method based on adaptive estimates of lower-order moments of the gradient. The stopping criterion was that the infinity norm of the expectation parameter matched the data expectation parameter to within 1 percent.

The results are shown in Fig. 5. The convergence to the final value can be monitored in terms of the normalized area under the optimization curve, $\sum_{t=1}^{T}(D_t - D_T)/(D_0 - D_T)$, where D_t is the divergence value at iteration t, and T is the final time. All methods achieved similar values of the divergence, except for the Euclidean gradient with non-orthogonal parametrization, which did not always reach the minimum. For the Fisher and Wasserstein gradients, the learning paths were virtually identical under the two different model parametrizations, as we already expected from the fact that these are covariant gradients. On the other hand, for the Euclidean gradient, the paths (and the number of iterations) were heavily dependent on the model parametrization, with the orthogonal basis usually being a much better choice than the non-orthogonal basis. In terms of the number of iterations until the convergence criterion was satisfied, the comparison is difficult because different methods work best with different step sizes. With the simple adaptive method and a suitable initial step size, the Wasserstein gradient was faster than the Euclidean and Fisher gradients. On the other hand, using Adam to adapt the step size, orthogonal Euclidean, Fisher, and Wasserstein were comparable.

6 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.

The Wasserstein distance has already been discussed with divergences in information geometry and also shown to be useful in machine learning, for instance in training restricted Boltzmann machines and generative adversarial networks. In this work, we used the Wasserstein distance to define a geometry on the parameter space of a statistical model. Following this geometry, we establish a corresponding natural gradient and displacement convexity on parameter space.

We presented an application of the Wasserstein natural gradient method to maximum likelihood estimation in hierarchical probability models. The experiments show that, in combination with a suitable step size, the Wasserstein gradient can be a competitive optimization method and even reduce the required number of parameter iterations compared both to Euclidean and Fisher gradient methods. It will be essential to conduct further experimental studies to better understand the effects of the learning rate, as well as the interplay of ground metric, model, and optimization problem. In our current implementation, the Wasserstein gradient involved heavier computational costs compared to the Euclidean and Fisher gradients. For applications, it will be important to explore efficient computation and approximation approaches.

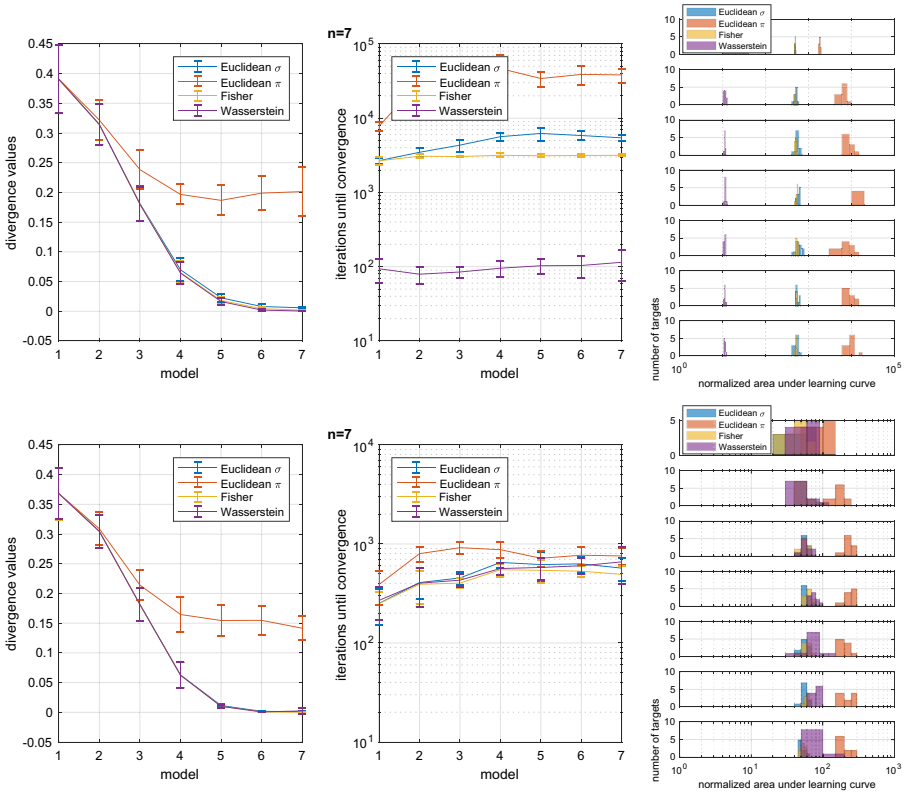

Fig. 5 Divergence minimization for random target distributions on $\{0,1\}^n, n = 7$, over k-interaction models with $k = 1, \ldots, n$. Shown is the average value of the divergence after optimization by Euclidean, Fisher, and Wasserstein gradient descent, and the corresponding number of gradient iterations. Orthogonal and non-orthogonal parametrization are indicated by σ and π. The right hand side shows histograms of the normalized area under the optimization curves. The top figures are using a simple adaptive method for selecting the step size, and the bottom figures are using Adam

Regarding the theory, 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 [32] 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?

The critical innovation coming from the Wasserstein gradient in comparison to the Fisher gradient is that it incorporates a ground metric in sample space. We suggest that this could have a positive effect not only concerning optimization, as discussed above, but also regarding generalization performance, in interplay with the optimization. The reason is that the ground metric on sample space provides means to introduce preferences in the hypothesis space. The specific form of such a regular-

ization still needs to be developed and investigated. In this regard, a natural question is how to define natural ground metric notions. These could be fixed in advance or trained.

We hope that this paper contributes to strengthening the emerging interactions between information geometry and optimal transport, in particular, to machine learning problems, and to develop better natural gradient methods.

Acknowledgements The authors would like to thank Prof. Luigi Malagò for his inspiring talk at UCLA in December 2017. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement n^{o} 757983).

Appendix

In this appendix we review the equivalence of static and dynamical formulations of the L^2-Wasserstein metric formally. For more details see [38].

Consider the duality of linear programming.

$$\begin{aligned}
&\frac{1}{2}W(\rho^0,\rho^1)^2\\
&\quad=\inf_{\pi\geq 0}\Big\{\int_\Omega\int_\Omega \frac{1}{2}d_\Omega(x,y)^2\pi(x,y)dxdy\colon \int_\Omega \pi dy=\rho^0(x),\ \int_\Omega \pi dx=\rho^1(y)\Big\}\\
&\quad=\sup_{\Phi^1,\Phi^0}\Big\{\int_\Omega \Phi^1(y)\rho^1(y)dy-\int_\Omega \Phi^0(x)\rho^0(x)dx\colon \Phi^1(y)-\Phi^1(x)\leq \frac{1}{2}d_\Omega(x,y)^2\Big\}.
\end{aligned}\tag{21}$$

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

$$\Phi^1(y)=\sup_{x\in\Omega}\ \Phi^0(x)+\frac{1}{2}d_\Omega(x,y)^2.\tag{22}$$

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

$$\frac{\partial\Phi(t,x)}{\partial t}+\frac{1}{2}g_x^\Omega(\nabla\Phi(t,x),\nabla\Phi(t,x))=0,\tag{23}$$

with $\Phi^0(x)=\Phi(0,x)$, $\Phi^1(x)=\Phi(1,x)$. Hence (21) becomes

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

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

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

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 (22) and Hamilton–Jacobi equation (23) 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.,

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

However, in a discrete sample space I, there is no path $\gamma(t)\in I$ connecting two discrete points. Thus the relation between (22) and (23) 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 [12,26].

Notations

We use the following notations.

Continuous/discrete sample space	Ω	I
Inner product	g^{Ω}	g^{I}
Gradient	∇	∇_G
divergence	div	div_G
Hessian in Ω	Hess	
Potential function set	$\mathcal{F}(\Omega)$	$\mathcal{F}(I)$
Weighted Laplacian operator	$-\nabla\cdot(\rho\nabla)$	$L(p)$

Continuous/discrete probability space	$\mathcal{P}_+(\Omega)$	$\mathcal{P}_+(I)$
Probability distribution	ρ	p
Tangent space	$T_\rho\mathcal{P}_+(\Omega)$	$T_p\mathcal{P}_+(I)$
Wasserstein metric tensor	g^W	g^W
Dual coordinates	$\Phi(x)$	$(\Phi_i)_{i=1}^n$
Primal coordinates	$\sigma(x)$	$(\sigma_i)_{i=1}^n$
First differential operator	δ_ρ	∇_p
Second differential operator	$\delta^2_{\rho\rho}$	
Gradient operator		∇_W
Hessian operator		Hess_W
Levi–Civita connection		$\nabla^W_{\cdot}\cdot$

Parameter space/Probability model	Θ	$p(\Theta)$
Inner product	g_θ	$g_{p(\theta)}$
Tangent space	$T_\theta\Theta$	$T_{p(\theta)}p(\Theta)$
L^2-Wasserstein matrix	$G(\theta)$	
L^2-Wasserstein distance	Dist	Dist
Second fundamental form		$B(\cdot,\cdot)$
Projection operator		H
Levi–Civita connection		$(\nabla^W_{\cdot}\cdot)^{\perp}$
Jacobi operator	J_θ	
First differential operator	∇_θ	
Gradient operator	∇_g	
Hessian operator	Hess_g	

References

1. Amari, S.: Neural learning in structured parameter spaces-natural Riemannian gradient. In: Mozer, M.C., Jordan, M.I., Petsche, T. (eds.) Advances in Neural Information Processing Systems 9, pp. 127–133. MIT, London (1997)
2. Amari, S.: Natural gradient works efficiently in learning. Neural Comput. **10**(2), 251–276 (1998)
3. Amari, S.: Information Geometry and Its Applications. Number volume 194 in Applied mathematical sciences. Springer, Tokyo (2016)
4. Amari, S., Karakida, R., Oizumi, M.: Information geometry connecting Wasserstein distance and Kullback-Leibler divergence via the Entropy-Relaxed Transportation Problem (2017). arXiv:1709.10219 [cs, math]
5. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN (2017). arXiv:1701.07875 [cs, stat]
6. Ay, N., Jost, J., Lê, H., Schwachhöfer, L.: Information Geometry Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics. Springer, Berlin (2017)
7. Bakry, D., Émery, M.: Diffusions hypercontractives. In: Azéma, J., Yor, M. (eds.) Séminaire de Probabilités XIX 1983/84, pp. 177–206. Springer, Berlin (1985)

8. Benamou, J.-D., Brenier, Y.: A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. Numerische Mathematik **84**(3), 375–393 (2000)
9. Campbell, L.: An extended **v**Cencov characterization of the information metric. Proc. Am. Math. Soc. **98**, 135–141 (1986)
10. Carlen, E.A., Gangbo, W.: Constrained Steepest Descent in the 2-Wasserstein Metric. Ann. Math. **157**(3), 807–846 (2003)
11. **v**Cencov, N.N.: Statistical Decision Rules and Optimal Inference. Translations of Mathematical Monographs, vol. 53. American Mathematical Society, Providence (1982). (Translation from the Russian edited by Lev J. Leifman)
12. Chow, S.-N., Huang, W., Li, Y., Zhou, H.: Fokker–Planck equations for a free energy functional or markov process on a graph. Arch. Ration. Mech. Anal. **203**(3), 969–1008 (2012)
13. Chow, S.-N., Li, W., Zhou, H.: A discrete Schrodinger equation via optimal transport on graphs (2017). arXiv:1705.07583 [math]
14. Chow, S.-N., Li, W., Zhou, H.: Entropy dissipation of Fokker–Planck equations on graphs. Discrete Contin. Dyn. Syst. A **38**(10), 4929–4950 (2018)
15. Chung, F. R. K.: Spectral Graph Theory. Number no. 92 in Regional conference series in mathematics. In: Published for the Conference Board of the mathematical sciences by the American Mathematical Society, Providence, R.I. (1997)
16. Frogner, C., Zhang, C., Mobahi, H., Araya-Polo, M., Poggio, T.: Learning with a Wasserstein loss (2015). arXiv:1506.05439 [cs, stat]
17. Gangbo, W., Li, W., Mou, C.: Geodesic of minimal length in the set of probability measures on graphs. accepted in ESAIM: COCV (2018)
18. Karakida, R., Amari, S.: Information geometry of wasserstein divergence. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information, pp. 119–126. Springer, Cham (2017)
19. Kingma, D. P., Adam, J. Ba.: A method for stochastic optimization (2014). CoRR, arXiv:1412.6980
20. Lafferty, J.D.: The density manifold and configuration space quantization. Trans. Am. Math. Soc. **305**(2), 699–741 (1988)
21. Lebanon, G.: Axiomatic geometry of conditional models. IEEE Trans. Inf. Theory **51**(4), 1283–1294 (2005)
22. Li, W.: Geometry of probability simplex via optimal transport (2018). arXiv:1803.06360 [math]
23. Li, W., Montufar, G.: Ricci curvature for parameter statistics via optimal transport (2018). arXiv:1807.07095
24. Li, W., Yin, P., Osher, S.: Computations of optimal transport distance with fisher information regularization. J. Sci. Comput. **75**, 1581–1595 (2017)
25. Lott, J.: Some geometric calculations on Wasserstein space. Commun. Math. Phys. **277**(2), 423–437 (2007)
26. Maas, J.: Gradient flows of the entropy for finite Markov chains. J. Funct. Anal. **261**(8), 2250–2292 (2011)
27. Malagò, L., Matteucci, M., Pistone, G.: 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, New York, NY, USA, 2011. ACM, pp. 230–242
28. Malagò, L., Pistone, G.: Natural gradient flow in the mixture geometry of a discrete exponential family. Entropy **17**(12), 4215–4254 (2015)
29. Mielke, A.: A gradient structure for reaction–diffusion systems and for energy-drift-diffusion systems. Nonlinearity **24**(4), 1329–1346 (2011)
30. Modin, K.: Geometry of matrix decompositions seen through optimal transport and information geometry. J. Geometr. Mech. **9**(3), 335–390 (2017)
31. Montavon, G., Müller, K.-R., Cuturi, M.: Wasserstein training of restricted boltzmann machines. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29, pp. 3718–3726. Curran Associates Inc, Red Hook (2016)
32. Montúfar, G., Rauh, J., Ay, N.: On the Fisher metric of conditional probability polytopes. Entropy **16**(6), 3207–3233 (2014)
33. Nelson, E.: Quantum Fluctuations. Princeton series in physics. Princeton University Press, Princeton (1985)
34. Otto, F.: The geometry of dissipative evolution equations: the porous medium equation. Commun. Partial Diff. Equ. **26**(1–2), 101–174 (2001)

35. Pascanu, R., Bengio, Y.: Revisiting natural gradient for deep networks. In: International Conference on Learning Representations 2014 (Conference Track) (2014)
36. Peters, J., Vijayakumar, S., Schaal, S.: Natural actor-critic. In: Gama, J., Camacho, R., Brazdil, P.B., Jorge, A.M., Torgo, L. (eds.) Machine Learning: ECML 2005, pp. 280–291. Springer, Berlin (2005)
37. Takatsu, A.: Wasserstein geometry of Gaussian measures. Osaka J. Math. **48**(4), 1005–1026 (2011)
38. Villani, C.: Optimal Transport: Old and New. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin (2009)
39. Wong, T.-K.: Logarithmic divergences from optimal transport and Rényi geometry (2017). arXiv:1712.03610 [cs, math, stat]
40. Yi, S., Wierstra, D., Schaul, T., Schmidhuber, J.: Stochastic search using the natural gradient. In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, New York, NY, USA. ACM, pp. 1161–1168 (2009)