# Optimal Transport on Riemannian Manifolds

Victor Armegioiu, *Student, TUM*

**Abstract**

The aim of this paper is to provide an overview of the machinery of differential geometry, Riemannian manifolds and their connection with Optimal Transport. We will examine the Optimal Transport problem in its typical formulations, characterize solutions and analyze them in the Riemannian setting. We will then showcase some specific application of Optimal Transport in Machine Learning.

**Index Terms**

Optimal transport, Riemannian geometry, Wasserstein space, style transfer, domain adaptation.

## I. Riemannian Geometry

THE purpose of this section is to provide the reader with some necessary tools for describing non-Euclidean geometries, specifically with a Riemannian structure, that shall be used further on to showcase results on Optimal Transport.

### A. Smooth Manifolds

**Remark A.1** The reader shall note that our treatment of smooth manifolds in this paper relies on the notion of **extrinsic** geometry. Specifically, we assume that the manifolds we analyze are embedded in some higher dimensional Euclidean space. This, however, need not be the case, since manifolds may well exist without being regarded as embedded in some higher dimensional manifold, and may be described as self-contained objects in a setting called 'intrinsic geometry'. We will, however, present our results, where possible, in the extrinsic setting. This is mainly for expository purposes, since being able to wander off the 'surface' of our manifolds into a higher dimensional Euclidean will have much stronger appeal to readers that are not yet familiar with Riemannian geometry. Furthermore, the Nash embedding theorem shows that it is always possible to embed arbitrary smooth manifolds into higher dimensional Euclidean spaces.

**Definition A.1** Let $U \subset \mathbb{R}^k$ and $V \subset \mathbb{R}^m$ be two open sets. A function $\phi : U \to V$ is called smooth iff it is infinitely differentiable, in that all its partial derivatives are continuous

$$d^N \phi = \frac{\partial^{n_1+n_2+...+n_k} \phi}{\partial x_1^{n_1} ... \partial x_k^{n_k}}$$

where $N = (n_1, ..., n_k) \in \mathbb{N}^k$.

If $\phi$ is bijective and both $\phi$ and $\phi^{-1}$ are smooth, then $\phi$ is called a **diffeomorphism**.

**Definition A.2** Let $M$ be an arbitrarily chosen set - An open set $U \subset M$ together with a diffeomorphism $\phi : U \to \phi(U) \subset R^m$ define a **coordinate chart** $(\phi, U)$ on $M$. Furthermore, two charts $(\phi_1, U_1), (\phi_2, U_2)$ are said to be smoothly compatible, if the composition map $\phi_2 \circ \phi_1^{-1} : \phi_1(U_1 \cap U_2) \to \phi_2(U_1 \cap U_2)$ is a diffeomorphism. A collection of charts $\{(\phi_i, U_i)\}_{i \in I}$, where $I$ is an index set, is called an atlas if $M \subset \cup_{i \in I} U_i$ and if all pairs of charts from the given collection are smoothly compatible.

### B. Tangent Spaces

Having equipped our manifolds with a smooth structure that allows us to identify them with subsets of $\mathbb{R}^m$ we will move toward defining the notions of tangent spaces, and derivatives.

**Definition B.1** Let $M \subset \mathbb{R}^k$ be a smooth $m$-dimensional manifold (unambiguously denoted as $m$-manifold henceforth), and fix a point $p$ on $M$. Let $\gamma : I \subset \mathbb{R}^k \to M$ be a smooth curve. $v \in \mathbb{R}^k$ is called a tangent vector if the following hold:

$$\gamma(0) = p \qquad \dot{\gamma}(0) = v$$

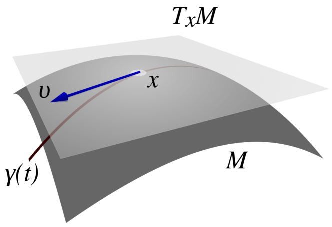
Victor Armegioiu is with the Department of Informatics, Technical University of Munich, Germany e-mail: victor.armegioiu@gmail.com.

Manuscript received Dec 11, 2020; revised Jan 11th, 2021.

That is, there is a smooth curve starting at $p$ with velocity $v$. It is trivial to show that we can always construct such a curve on any given manifold. The equivalence class of all such curves, is called the **tangent space** of $M$ at $p$, denoted as $T_pM$

$$T_pM = \{v | \gamma(0) = p, \dot{\gamma}(0) = v\}$$

Fig. 1: Tangent space at $x$ visualization



Note that we defined the tangent space at a point as a subspace of $\mathbb{R}^k$. This typically means that the plane need not actually intersect the manifold, and Figure 1 actually displays the tangent space shifted for illustration purposes, hence we're actually looking at $T_xM + x$.

**Definition B.2** Let $M \subset \mathbb{R}^k$ be a smooth $m$-dimensional manifold, and a smooth function $f : M \to \mathbb{R}^d$. Fixing a point $p$ on $M$. The **derivative** of $f$ at $p$ is the Jacobian map

$$df(p) : T_pM \to \mathbb{R}^d$$

defined as follows: Construct a curve $\gamma : \mathbb{R} \to M$ with initial conditions $\gamma(0) = p, \dot{\gamma}(0) = v$. As noted before, this is always possible. Now, we define the directional derivative $df(p)v \in \mathbb{R}^d$

$$df(p)v := \frac{d}{dt}\bigg|_{t=0} f(\gamma(t)) = \lim_{h \to 0} \frac{f(\gamma(h)) - f(p)}{h}$$

Having defined the concept of derivatives and tangent spaces, we move on to understanding how to actually form a basis of the tangent space, using coordinate charts. This will prove quite useful later on, particularly when defining Riemannian metrics.

**Theorem B.1** Let $M \subset \mathbb{R}^k$ be a smooth $m$-manifold. Fix a point $p$ on $M$, and choose a coordinate chart $(\phi_0, U_0)$ such that $p \in U_0 \subset M$, and $\phi_0 : U_0 \to \Omega_0 \in \mathbb{R}^m$. Denote $\psi_0 := \phi_0^{-1} : \Omega_0 \to U_0$. Let $x_0 := \phi_0(p)$, then the following holds:

$$T_pM = \text{im } (d\psi_0(x_0) : \mathbb{R}^m \to \mathbb{R}^k) \tag{1}$$

*Proof.* An inclusion argument is trivial to make, i.e. for a small enough $t$ and $\xi \in \mathbb{R}^m$, such that $x_0 + t\xi \in \Omega_0$, we may construct a curve $\gamma : (-\epsilon, \epsilon) \to M$, such that $\gamma(t) := \psi_0(x_0 + t\xi)$. Naturally, we get:

$$\gamma(0) = p, \ \dot{\gamma}(0) = \frac{d}{dt}\bigg|_{t=0} \psi_0(x_0 + t\xi) = d\psi_0(x_0)\xi$$

This shows that, indeed $d\psi_0\xi \in T_pM$. To show that $d\psi_0$ covers all of $T_pM$, we note that it is always possible to create a smooth extension of $\phi_0$, denoted as $\Phi : U \subset \mathbb{R}^k \to \mathbb{R}^m$ such that $\Phi$ agrees with $\phi_0$ on $U \cap M$, namely $\Phi(U \cap M) = \phi(U_0)$. Furthermore, it follows that $\Phi(\psi_0(x)) = \phi_0(\psi_0(x)) = x$, hence we get $d\Phi(\psi_0(x))d\psi_0(x) = \text{id } : \mathbb{R}^m \to \mathbb{R}^m$. This shows that the map $d\psi_0(x)$ has a left inverse, which makes it an injective map, implying $\ker d\psi_0(x) = \{0\}$. Since, via the rank-nullity theorem we have, $\dim \text{im } d\psi_0(x_0) = m - \dim \ker d\psi_0(x_0) = m$. This completes the proof. □

Note, however, that this is one of many equivalent formulations for the tangent space $T_pM$. However, this one strikes a great balance between its usefulness and ease of understanding, and will be sufficient for our purposes.

### *C. Covariant Derivative*

So far, all of our efforts have been concentrated on finding identifications of smooth manifolds to subspaces of $\mathbb{R}^k$, in order to transfer our notions of analysis of flat surfaces to arbitrarily shaped ones. Starting with this section, we will make use of these identifications with Euclidean spaces to get a sense of how to 'move' on our manifolds, how to monitor changes along paths that lie on curved surfaces, and how to actually measure vectors and curvature.

**Definition C.1** Let $M \in \mathbb{R}^k$ be a smooth $m$-manifold. We define a **vector field** as a smooth map, $X : M \to \mathbb{R}^k$, such that for any $p \in M$, $X(p) \in T_pM$. We denote a **vector field along a curve** $\gamma : I \subset \mathbb{R} \to M$ to be a smooth map $X : I \to \mathbb{R}^k$, which satisfies $X(t) \in T_{\gamma(t)}M$, $\forall t \in I$.

Taking the derivative of $X(t)$ brings us to the many challenges of working on manifolds - namely, the resulting vector field $\dot{X}(t)$ is not bound to be tangent to $M$, so we need some projection mechanism to split the resulting field into two orthogonal components, one of which is tangent to the manifold, while the other is normal.

**Theorem 3.1** Let $M \subset \mathbb{R}^k$ be a smooth $m$-manifold, choose a chart $(\phi : U \to \Omega \subset \mathbb{R}^m, U)$ and fix an arbitrary point $p \in U$. Denote the inverse map $\psi := \phi^{-1} : \Omega \to U$. Then, the orthogonal projection $\Pi = \Pi^2 = \Pi^T$, such that $\Pi(p)v = v \iff v \in T_pM$, is uniquely induced by the inner product structure inherited from the ambient space, and has the form $\Pi(p) = d\psi(\phi(p))\big(d\psi(\phi(p))^T d\psi(\phi(p))\big)^{-1} d\psi(\phi(p))^T$.

*Proof sketch.* Recall from **Theorem B.1** that the tangent space $T_pM$ is spanned by the columns of $d\psi(\phi(p))$. Considering that the inner product inherited from the ambient space induces a norm $(\mathbb{R}^k, \|\cdot\| := \sqrt{\langle\cdot,\cdot\rangle})$, we can uniquely identify the projection of an arbitrary vector $p$ onto $T_pM$ as

$$\text{proj}_{T_pM}(v) = \underset{w \in \text{ im } d\psi(\phi(p))}{\arg\min} \|w - v\|$$

which has the known exact solution $w^* := \big(d\psi(\phi(p))^T d\psi(\phi(p))\big)^{-1} d\psi(\phi(p))^T v$, easily obtainable via solving the Ordinary Least Squares (OLS) problem (in the $\|\cdot\|_{L_2}$ norm). Naturally, the actual projection can be recovered as $\Pi(p) = d\psi(\phi(p))\big(d\psi(\phi(p))^T d\psi(\phi(p))\big)^{-1} d\psi(\phi(p))^T$. □

Given **Theorem C.1** we can now write the decomposed form of the derivative $\dot{X}(t)$ as

$$\dot{X}(t) = \Pi(\gamma(t))\dot{X}(t) + (\mathbb{1} - \Pi(\gamma(t)))\dot{X}(t)$$

**Definition C.2** Let $\gamma : I \subset \mathbb{R} \to M$ be a smooth curve, and consider any smooth vector field $X$, with $X(t) \in T_pM_{\gamma(t)}$. The **covariant derivative** of $X$ along $\gamma$ is the vector field defined as

$$\nabla X := \Pi(\gamma)\dot{X},\ \nabla X \in T_\gamma M$$

### *D. Riemannian Metrics and Curvature*

**Definition E.1** Let $M \subset \mathbb{R}^k$ be a smooth $m$-manifold. For every $p \in M$ we denote the tangent bundle of $M$ by $TM := \cup_p(\{p\} \times T_pM)$. On every tangent space $T_pM$, we assume a smoothly varying symmetric positive-definite bilinear form $g : T_pM \times T_pM$, $g$ is called a **Riemannian metric**, inducing a metric structure on $M$, hence $(M, g)$ is called a **Riemannian manifold**.

A Riemannian metric induces a scalar product and a norm on the tangent bundle, so for any $u, v \in T_pM$ we have $\langle u, v\rangle := g(u, v)$ and the norm is defined in the usual sense as $\|u\| = \sqrt{g(u,u)}$. Working with coordinates, let $(\phi, U)$ denote a chart on $M$, with inverse $\psi := \phi^{-1} : \Omega : U$. Fix a point $p \in U$, and choose arbitrary vectors $u, v \in T_pM$. Since $T_pM = \text{im } d\psi(\phi(p))$, we can write any $u \in T_pM$ as a linear combination of basis vectors, hence:

$$g(u, v) = \sum_{i,j=1}^{m} \Big\langle \frac{\partial \psi}{\partial x_i}(\phi(p)), \frac{\partial \psi}{\partial x_j}(\phi(p)) \Big\rangle u_i v_j \tag{2}$$

**Definition E.2** Let $M \subset \mathbb{R}^k$ be a smooth $m$-manifold, equipped with a Riemannian structure. Let $\gamma : \mathbb{R}^2 \to M$ be a smooth curve on $M$, parameterized by $(s, t)$. Let $X : \mathbb{R}^2 \to T_{\gamma(s,t)}M$ be a smooth vector field along $\gamma$. The partial covariant derivatives of $X$ with respect to its input coordinates, are defined as:

$$\nabla_s X = \Pi(\gamma)\frac{\partial X}{\partial s},\ \ \nabla_t X = \Pi(\gamma)\frac{\partial X}{\partial t} \tag{3}$$

Let $p \in M$, then **Riemann curvature tensor** associates $p$ with the bilinear map $R_p : Tp_M \times Tp_M \to \mathcal{L}(Tp_M, Tp_M)$ defined as

$$R_p(u, v)w = (\nabla_s \nabla_t X - \nabla_s \nabla_s X)(0, 0) \tag{4}$$

where $u, v, w \in T_pM$ and $X$ is a smooth vector field along a curve, as previously defined. We note the following initial conditions:

$$\gamma(0,0)=p,\ \partial_s\gamma(0,0)=u,\ \partial_t\gamma(0,0)=v,\ X(0,0)=w.$$

Intuitively $R_p$ measures the commutativity of the partial covariant derivatives, to establish whether the space is locally isometric to Euclidean space, or flat. The rationale here, is that, in flat space, it should not matter in which order we monitor changes along the axes - Recall that partial differentiation in $\mathbb{R}^d$ is always commutative, via Schwarz's theorem $\partial_x\partial_y = \partial_y\partial_x$. See Figure 2 for an illustration of different types of curvature.

### *E. Geodesics and the Exponential Map*

In order to generalize the idea of a straight line as a shortest path between two points to arbitrarily curved Riemannian manifolds, we turn to geodesics. Formally, given a Riemannian manifold $(M,g)$, two points $p,q \in M$ and a smooth parameterized curve $\gamma : [0,1] \rightarrow M$, with $\gamma(0) := p, \gamma(1) := q$ the length of the curve connecting $p,q$ can be simply characterized as the functional

$$L(\gamma) = \lim_{n\rightarrow\infty}\sum_{i=1}^{n}\|\gamma(t_i)-\gamma(t_{i-1})\| = \int_0^1 \|\dot{\gamma(t)}\|dt = \int_0^1 \sqrt{g\big(\dot{\gamma(t)},\dot{\gamma(t)}\big)}dt \tag{5}$$

**Definition E.1** Let $M \subset R^k$ be a smooth $m$-manifold and $\gamma : I \subset \mathbb{R} \rightarrow \mathbb{M}$ be a smooth curve on $M$. $\gamma$ is called a **geodesic** if it is a minimizer of the length functional $L(\gamma)$. Consequently, the distance between two points $p,q \in M$ can be defined as $d(p,q) := \inf_\gamma L(\gamma)$, where the norm is taken with respect the given metric tensor $g$.

In keeping with Remark A.1, we note that we have made extensive use of the ambient space in our definitions, particularly when defining tangent vectors as initial speeds of curves. As a consequence, we need some Riemannian apparatus to bring information from the tangent space back on the manifold.

**Definition E.2** Let $M$ be a smooth $m$-manifold, and $\gamma : [0,1] \rightarrow M$ be the unique constant speed geodesic characterised by initial data $\gamma(0) = p, \dot{\gamma(0)} = v$. The exponential map is defined as $\exp_p(v) := \gamma(1)$, which assigns to any tangent vector in $T_pM$ the endpoint of the geodesic on $M$.

The exponential map also shares a deep connection with geodesics, given that geodesic curves, initialized with $\gamma(0) = p, \dot{\gamma(0)} = v$, can be written as $\gamma(t) = \exp_p(tv)$.

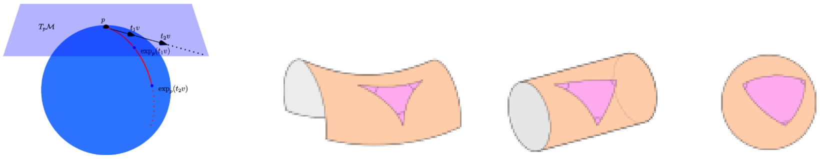


Fig. 2: **Left side** shows that for $v \in T_pM$, the exponential map brings contractions/dilations $tv$ of $v$ to the same geodesic. **Right side** depitcts the effective of negative, flat and positive curvature on **geodesic** triangles. Source: http://www.science4all.org

## II. Optimal Transport Problem

The optimal transport problem is a mathematical framework which formalizes the study of resource allocation and transport plans. Specifically, the field's central problem is finding transportation schemes that show how to transfer some arbitrary mass from some place to another, while minimizing some predefined cost function.
Let $\mathcal{X}, \mathcal{Y}$ be Radon spaces, with the Radon measures $\mu, \nu$ with $\text{supp}(\mu) \subset \mathcal{X}, \text{supp}(\nu) \subset \mathcal{Y}$ and a cost function $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$. We distinguish the following formulations of the optimal transport problem

- **Monge formulation** In this formulation, we are seeking a transport plan $T : \mathcal{X} \rightarrow \mathcal{Y}$, which minimizes some convex cost function $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$. This is called the Monge problem, and may be formally written as
$$\inf_{T_{\#\mu}=\nu} \int_{\mathcal{X}} c(x,T(x))d\mu(x) \tag{6}$$
Where the infimum is taken over all measurable maps $T$, such that the pushforward $T_{\#\mu}(M) := \mu(T^{-1}(M))$ agrees with $\nu$. Naturally, $T_{\#\mu}$ induces a Borel probability measure as well.
- **Kantorovich formulation** The Monge formulation describes the optimal transport map $T$ in a very strict sense, such that it may very well be impossible to find an optimal map $T$ depending on the measures $\mu, \nu$. For example, there is no

measurable mapping that can transport a Dirac measure $\mu$ ($\delta_x(A) = 1$ iff $x \in A$) to a non-Dirac measure $\nu$. Hence, the Kantorovich formulation approaches the problem from a probabilistic perspective

$$\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\gamma(x,y) \tag{7}$$

where $\gamma$ is a joint measure in the collection of all possible couplings $\Gamma(\mu,\nu)$ which have marginals $\mu, \nu$.

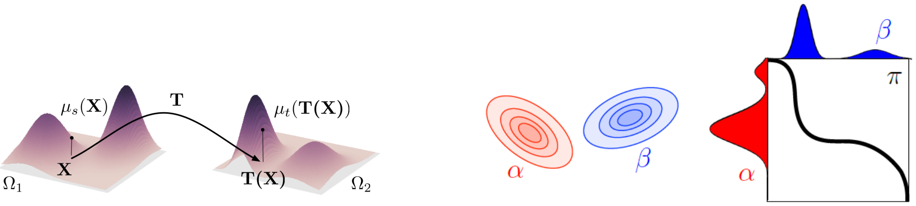


Fig. 3: The **left** figure is an illustration of the **Monge** problem, where an optimal transport map $T$ transfers density from $\Omega_1$ to $\Omega_2$, while enforcing density convservation. The **right** figure illustrates the **Kantorovich** reformulation, where as $\pi$ describes the optimal transport plan. Heuristically speaking, $\pi$ describes a symmetric infinite dimensional bistochastic matrix where $\pi(x,y) :=$ how much mass goes from $x$ to $y$ (or vice-versa) in the optimal solution.

Both formulations describe a minimization scheme based on cost. However, there is another particularly enlightening point of view, which reframes the transport problem as a profit maximization scheme. In order to do that, we will first have to define some useful terms.

### *A. Kantorovich problem and its dual*

**Definition A.1** Let $\mathcal{X}, \mathcal{Y}$ be arbitrary sets, and $\Gamma \subset \mathcal{X} \times \mathcal{Y}$ a subset of their cartesian product. If for a given cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the following holds for any $n$ arbitrarily chosen points $\{(x_i, y_i)\}|_{i=1}^{n}$

$$\sum_{i=1}^{n} c(x_i, y_i) \leq \sum_{i=1}^{n} c(x_i, y_{i+1}) \tag{8}$$

with $y_{n+1} := y$ then $\Gamma$ is said to be **c-cyclically monotone**. Furthermore, a transfer plan concentrated on $\Gamma$ is also said to be c-cyclically monotone. We say that a measure $\mu$ is concentrated on some $\Theta \in \mathcal{X}$ if $\mu(\mathcal{X} \setminus \Theta) = 0$.

The notion of cyclical monotonicity allows us to construct a greedy procedure for optimizing transport plans. Given a transport plan, choose an arbitrary starting point $x_1$ which is bound to send some measurable quantity to $y_1$, as assigned in this initial plan. Take some portion of the mass sent by $x_1$ to $y_1$ and redirect it to some $y_2$ that is closer to $x_1$ under the topology induced by the cost function $c(x,y)$. Doing this, generates a lowering of cost of $c(x_1,y_1) - c(x_1,y_2)$. Naturally, given the extra resources now at $y_2$, $x_2$ will need to reroute some of its mass to some closer $y_3$, so on, and so forth, until some $x_N$ closes the loop by sending some mass to $y_{n+1} := y$. The new plan is strictly better if it is c-cyclically monotone, as per equation 8. This also suggests an optimization procedure, that is similar in nature to minimum cost max flow on bipartite graphs - here, we can systematically exploit cycles, looking for augmentations to our costs. Once this cannot be done anymore, it is very likely that our transport plan is optimal.

Switching from minimizing costs, to maximizing profit, the **Kantorovich dual problem** is stated as follows

$$\sup \left\{ \int_{\mathcal{Y}} \phi(y) d\nu(y) - \int_{\mathcal{X}} \psi(x) d\nu(x) \middle| \forall x, y \;\; \phi(y) \leq \psi(x) + c(x,y) \right\} \tag{9}$$

Which can be interpreted in a practical fashion as buying some unit of mass at price $\psi(x)$ and then selling at $\phi(y)$. Naturally, in order to maximize profit and also maintain competitiveness

$$\forall x, y \;\; \phi(y) \leq \psi(x) + c(x,y) \tag{10}$$

has to hold. Otherwise, if the selling cost $\phi(y) > \psi(x) + c(x,y)$ the buyer can just handle their own transport cost, plus buying the materials at the price $\psi(x)$ and get a better deal, hence the price $\phi(y)$ would not be competitive.
In order to characterize solutions of this problem, we introduce the notions of **c-convex** functions and **c-subdifferentials**.

**Definition A.2** Let $\mathcal{X}, \mathcal{Y}$ be arbitrary sets, and a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. A function $\Psi : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is said to be c-convex if

$$\forall x \in \mathcal{X},\ \psi(x) = \sup_{y \in \mathcal{Y}} \{\psi^c(y) - c(x,y)\} \tag{11}$$

such that

$$\forall y \in \mathcal{Y},\ \psi^c(y) = \inf_{x \in \mathcal{X}} \{\psi(x) + c(x,y)\} \tag{12}$$

For a given c-convex function, its c-subdifferential at $x$ is denoted as $\partial_c \psi(x) := \{y \in \mathcal{Y} | \psi(x) = \psi^c(y) - c(x,y)\}$. Moreover, a function $\psi$ is called $\frac{d^2}{2}$-convex iff $\psi(x) + \frac{|x|^2}{2}$ is convex in the Euclidean sense.

**Remark A.1** Given measure space $(\mathcal{X}, \mu)$ - with the implied power set $\sigma$-algebra for simplicity - we shall say that a property on $\mathcal{X}$ holds $\mu$-**almost everywhere**, if the reunion of all subsets of $\mathcal{X}$ upon which this property is violated (denoted as $\Omega \subset \mathcal{X}$) has measure 0, that is $\mu(\Omega) = 0$.

**Proposition A.1** The dual reformulation

$$\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\gamma(x,y) = \sup_{(\phi,\psi), \phi - \psi \le c} \int_{\mathcal{Y}} \phi(y) d\nu(y) - \int_{\mathcal{X}} \psi(x) d\nu(x) = \sup_{\psi} \int_{\mathcal{Y}} \psi^c(y) d\nu(y) - \int_{\mathcal{X}} \psi(x) d\nu(x) \tag{13}$$

where $\psi, \phi$ are integrable functions with respect to the measures $\mu, \nu$, is shown by Villani [1][chapter 5] to lead to the following statements, describing a solution to the Kantorovich problem.

If $c$ is real-valued, and the optimal cost $C(\mu,\nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int c\, d\gamma$ is finite, then there is a measureble c-cyclically monotone set $\Theta \in \mathcal{X} \times \mathcal{Y}$, such that for any $\gamma \in \Gamma(\mu,\nu)$, the following are equivalent

1) $\gamma$ is optimal;
2) $\gamma$ is c-cyclically monotone;
3) There is a c-convex $\psi$ such that, $\gamma$-almost surely, $\psi^c(y) - \psi(x) = c(x,y)$;
4) There exist $\psi : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ and $\phi : \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$, such that $\phi(y) - \psi(x) \le c(x,y), \forall (x,y)$ with equality $\gamma$-almost surely;
5) $\gamma$ is concentrated on $\Theta$.

Furthermore, if $c$ is real valued and the cost $C(\mu,\nu) < \infty$, and we have finite upper bounds on all $c(x,y)$, then both the primal and dual Kantorovich have solutions, hence in the latter expressions we can also impose that $\psi$ be c-convex and $\phi = \psi^c$. In addition, if some light conditions are met [1] [part I, chapter 5], there is a closed c-cyclically monotone set $\Theta \subset \mathcal{X} \times \mathcal{Y}$, such that for any $\gamma \in \Gamma(\mu,\nu)$ and for any c-convex $\mu$-integrable function:

1) $\gamma$ is optimal in the Kantorovich **primal** problem iff $\gamma(\Theta) = 1$, that is $\gamma$ is concentrated on $\Theta$;
2) $\psi$ is optimal in the Kantorovich **dual** problem if $\Theta \subset \partial_c \psi$.

### *B. Solving the Monge-Kantorovich Problem on Riemannian Manifolds*

Let $(M, g)$ be a Riemannian manifold, where $g$ denotes the metric tensor which gives an inner product structure on the tangent spaces. Let $c$ denote the geodesic distance, and $\mathcal{P}(M)$ the space of probability measures on $M$. The **Monge problem with quadratic** cost, can be formalized as follows, given $\mu, \nu \in \mathcal{P}(M)$, as the minimization problem

$$\inf_{T_{\#\mu} = \nu} \int_M c(x, T(x))^2 d\mu(x) \tag{14}$$

We recall that $T_{\#\mu}(A \subset M) := \mu(T^{-1}(A \subset M))$ denotes the pushforward of $\mu$ by $T$, which has to agree with $\nu$ as part of the constraints.

Due to McCann [2], we have a generalization of Brenier's polarization theorem [3][chapter 3] to compact Riemannian manifolds, with more general cost functions. We follow the exposition of Figalli and Villani [4], for the presentation of the following results.

**Theorem B.1** Let $(M, g)$ be a compact Riemannian $m$-manifold. Construct $m$-dimensional Hausdorff measures (**Riemannian volume**, this is the Riemannian analogue to Lebesgue measures in Euclidean space) as $\text{vol}(dx) = \sqrt{\det(g) dx_1 \cdot \ldots \cdot dx_m}$. Let $\mu = f(x) d\text{vol}(dx), \nu = g(y) d\text{vol}(dy), \mu, \nu \in \mathcal{P}(M)$, be probability measures on $M$, and consider the quadratic cost $\frac{1}{2} c(x,y)^2$. Then, the following hold

1) The Monge problem has a unique solution $T$;
2) The map $T$ can be written as $T(x) = \exp_x(\nabla \psi(x))$ for a $\frac{d^2}{2}$-convex function $\psi : M \to \mathbb{R}$;
3) Form $\mu$-almost all $x$, there exists a **unique** minimizing geodesic from $x$ to $T(x)$, which is given by the map $t \mapsto \exp_x(t \nabla \psi(x))$;

4) The Jacobian has the form $J_x T = \frac{f(x)}{g(T(x))}$ $\mu$-almost everywhere.

Consequences $1), 2)$ of **Theorem B.1** characterize the uniqueness and analytic form of the solution. Consequence $3)$ gives us a fantastic insight, which was not developed until recently. Specifically, this allows one to perform interpolations across unique minimizing curves that connect arbitrary points on a given manifold $M$. This has practical advantages even in applied Machine Learning - the ability to describe a way to move across shortest paths between points opens up possibilities for interpolation, Riemannian adaptive optimizers (think about a Riemannian version of Stochastic Gradient Descent). Consequences $2), 3)$ and $4)$ allow us to study curvature information along geodesics on $M$, and will provide a way to directly measure the smoothness and continuity of the optimal map $T$.

For a proof sketch of these results, we refer the reader to Figalli and Villani [4], which can mainly be formulated as a combination of the statements listed under Proposition A.1. We note that optimal transport maps $T$ may not be smooth, or even continuous, in the Riemannian setting. This might happen due to several obstructions, such as negative sectional curvature. Indeed, one could take the example of $\mathbb{H}^2$, the hyperbolic plane and consider a triangle $\Delta OAB$, where $O$ denotes the origin of the plane. Due to negative curvature (as illustrated in 2), Pythagora's theorem degenerates into an inequality, whereas we can obtain $d(O,A)^2 + d(O,B)^2 < d(A,B)^2$. This would contradict the cyclical monotonicty required for the support of an optimal transport plan, as shown in Proposition A.1, items $1)$ and $2)$, since one would be able to infinitely improve on any given plan, given the violation of Pythagora's theorem.

### *C. Computational Applications of Optimal Transport*

*1) Wasserstein Space:* Let $\mathcal{X} \subset \mathbb{R}^d$ be a separable, Banach space, we define the Wasserstein space as

$$\mathcal{W}(\mathcal{X}) = \left\{ \mu \middle| \int_{\mathcal{X}} \|x\| d\mu(x) < \infty \right\} \tag{15}$$

Given arbitrary Borel probability measures with finite second moment $\mu, \nu \in \mathcal{P}(\mathcal{X})$ the Kantorovich optimal transportation cost (here $c(x,y) = \|x-y\|^p$) introduces a natural metrization over the Wasserstein $p$-space as

$$W_p(\mu,\nu) = \left( \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} \|x-y\|^p d\gamma(x,y) \right)^{\frac{1}{p}} \tag{16}$$

Showing the triangle inequality is the only metric axiom that is non-trivial to show, for which we refer the reader to Villani [1] [chapter 6].
Due to Otto [5], the Wasserstein space $\mathcal{W}$ can be endowed with an infinite-dimensional Riemannian metric tensor $g^{\mathcal{W}(\mathcal{X})}$, in a formal sense, sometimes referred to as a density manifold $(\mathcal{W}(\mathcal{X}), g^{\mathcal{W}(\mathcal{X})})$. Having established the Riemannian structure of $\mathcal{W}(\mathcal{X})$, we employ Theorem B.1, proposition 3) to establish geodesics between Borel probability measures $\mu, \nu \in \mathcal{W}(\mathcal{X})$, which allows for interpolation between arbitrary measures supported on $\mathcal{X}$.

*2) Wasserstein Style Transfer:* The work by Mroueh [6] is mainly based on McCann's insight regarding interpolation in the Wasserstein space. Precisely, Theorem B.1, item 3), coupled with the Riemannian structure of $\mathcal{W}$ allows us to perform interpolations between arbitrary finite second momemnt Borel probability measures $\mu, \nu \in \mathcal{W}$.

*Wasserstein Style Transfer* proposes a novel optimal transport procedure for taking a content image $I_c$, along with a target style image $I_s$ and then produce a new image $\tilde{I_s}$ where the content image is redrawn in the style of the style image.

**Methodology**

Assume we are working with images $I \in \mathbb{R}^{C\times(W\times H)}$, where $C$ denotes the number of channels, and $H, W$ denote the height, and width respectively. Given a set of feature extractors $\{F_j\}|_{j=1}^n : \mathbb{R}^{C\times(W\times H)} \to \mathbb{R}^d$ which map images to latent codes, we can define an encoder mapping $E : \mathbb{R}^{C\times(W\times H)} \to \mathcal{W}(\mathbb{R}^d)$, $I \to \nu_I = \frac{1}{n}\sum_{j=1}^n \delta_{F_j(I)}$ which takes an input image, and maps it to an **empirical** measure over its latent representations. It is implicitly assumed that this map is invertible, hence there exists a decoder $D : \mathcal{W}(\mathbb{R}^d) \to \mathbb{R}^{C\times(W\times H)}$ such that $D(E(I)) = I$.

Let $I_c, I_s$ denote the content and style images, respectively. The author proposes an encoder decoder architecture, which take the images, encodes both of them as Gaussian empirical measures - where the parameters $\mu_c = \frac{1}{n}\sum_{j=1}^n F_j(I_c)$, $\Sigma_c = \frac{1}{n}\sum_{j=1}^n (F_j(I_c) - \mu_c)(F_j(I_c) - \mu_c)^T$ fully specificy the empirical distribution $\nu_c$ over the content image $I_c$. The style image $I_s$ is encoded to a Gaussian distribution $\nu_s$ in the exact same fashion. An optimal interpolant $\nu$ between $\nu_c$, $\nu_s$ has to be found based on a tradeof between style and content, and finally decode the stylized image using the decoder, $\tilde{I_s} = D(\nu)$.

Therefore, in order to have full control over content/style preservation, a parameter $t \in [0, 1]$ is introduced, and the optimal interpolant $\nu$ is found as a solution to the minimization problem

$$\min_{\nu}(1-t)W_2^2(\nu_c, \nu) + tW_2^2(\nu_s, \nu) \tag{17}$$

Where $W_2^2$ denotes the squared Wasserstein distance, with $p = 2$. Fortunately, the optimal solution $\nu$ has closed form when $\nu_c, \nu_s$ are Gaussian distributions, as they are in our case.

*3) Optimal Transport for Domain Adaptation:* The work of Flamary et. al [7] proposes a novel approach for the task of domain adaptation, whereas the goal is to produce domain invariant models, which can generalize well beyond their training distribution. The authors study the general case of classification tasks, where it is often the case that one might have access to several different datasets, out of which only a small proportion have labels. In order to effectively make use of all the data, their solution involves finding an optimal transport plan between two given related datasets $\Omega_s, \Omega_t$, where $\Omega_s$ (the source domain) is labeled, and used for training while $\Omega_t$ (the target domain) is unlabeled, and is **not** used for training. We note that a crucial assumption to make is that the conditional distribution of the labels, conditioned on the domains should be the same, which means given an optimal transport map $T : \Omega_s \to \Omega_t$, $P(y|x \in \Omega_s) = P(y|T(x) \in \Omega_t)$ holds.

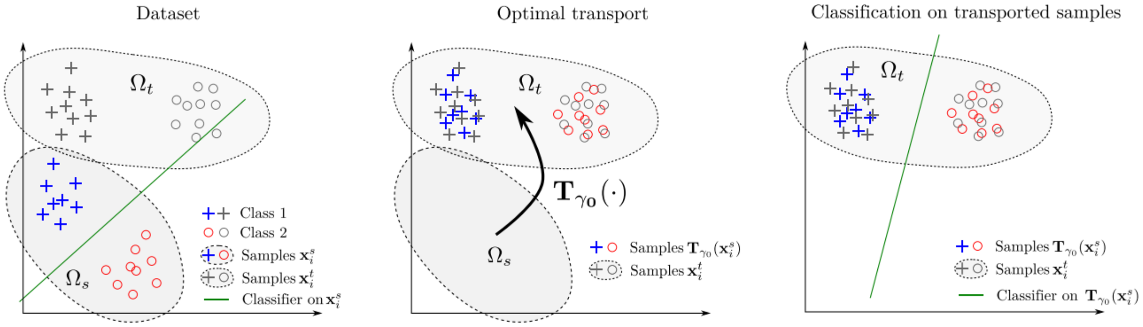

Dataset
$\Omega_t$
$\Omega_s$
Class 1
Class 2
Samples $\mathbf{x}_i^s$
Samples $\mathbf{x}_i^t$
Classifier on $\mathbf{x}_i^s$
Optimal transport
$\Omega_t$
$\mathbf{T}_{\gamma_0}(\cdot)$
$\Omega_s$
Samples $\mathbf{T}_{\gamma_0}(\mathbf{x}_i^s)$
Samples $\mathbf{x}_i^t$
Classification on transported samples
$\Omega_t$
Samples $\mathbf{T}_{\gamma_0}(\mathbf{x}_i^s)$
Samples $\mathbf{x}_i^t$
Classifier on $\mathbf{T}_{\gamma_0}(\mathbf{x}_i^s)$

Fig. 4: Visualization of the application of an optimal transport mapping $T$, mapping the source domain to the target domain. Source: Flamary et al [7]

The authors use the Kantorovich relaxation for finding an optimal transport plan $\gamma_0 \in \mathcal{P} = \{\gamma \in \mathbb{R}_+^{n_s \times n_t} | \gamma 1_{nt} = \mu_s, \gamma^T 1_{ns} = \mu_t\}$ - which reads as finding an optimal bistochastic coupling $\gamma_0$ with marginals $\mu_s$ and $\mu_t$, respectively where $\mu_s = \frac{1}{n_s}\sum_{i=1}^{n_s}\delta_{x_i^s}$, $\mu_t = \frac{1}{n_t}\sum_{i=1}^{n_t}\delta_{x_i^t}$ are the empirical measures over the samples from the source and target domains. Given a convex cost function $c : \Omega_s \times \Omega_t \to \mathbb{R}^+ \cup \{\infty\}$, we define the cost matrix $C_{i,j} := c(x_s^i, x_t^j)$ and naturally redefine the Kantorovich problem in the discrete sense as

$$\inf_{\gamma \in \mathcal{P}} \langle \gamma, C \rangle_F + \lambda \Omega_s(\gamma) + \eta \Omega_t(\gamma) \tag{18}$$

Where $\langle A, B \rangle_F := Tr(A^T B)$ is the Frobenius inner product, and $\Omega_s(\gamma)$, $\Omega_t(\gamma)$ are domain regularization terms, whose contributions are controlled by the scalars $\lambda, \eta$. $\Omega_s(\gamma) := \sum_{i,j}\gamma(i,j)\log\gamma(i,j)$ denotes the negative entropy of the optimal transport plan $\gamma$. The purpose of this term is to smoothen $\gamma$, by maximizing its entropy - without this restriction, the transport plan would almost surely become sparse, and less interpretable, since it would place 0 mass transfer on most pairs, i.e. $\gamma(x_i^s, x_i^t)$ = 0.

The term $\Omega_t(\gamma) := \frac{1}{n_s^2}\sum_{i,j=1} S(x_i^s, x_j^s)\|\hat{x_i^s} - \hat{x_j^s}\|_2^2$, where $\hat{x_i^s}$ is where the source point $x_i^s$ is transported into the target domain $\Omega_t$, and the function $S : \Omega_s \times \Omega_s \to \mathbb{R}^+$ denotes the similarity between source points $x_i^s, x_j^s$, ensures that similar samples in the source domain, are also similar after transport in the output domain.

## III. Conclusion

We have seen that optimal transport provides a particularly versatile mathematical framework that can be successfully applied to completely unrelated tasks. Furthermore, with the recent developments in connecting Riemannian geometry and optimal transport, this has enabled even richer applications, as we have noted in the case of Borel probability measures interpolation. The field of computational optimal transport is becoming further and further developed, particularly through the work of Marco Cuturi [8].

## Acknowledgment

The author would like to thank Severin Reiz for the opportunity to work on this subject, even though it was not listed on the topic list. The author would also like to thank Felix Dietrich for advice on the applied side of the literature.

# References

[1] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[2] Robert J McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

[3] Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[4] Alessio Figalli and Cédric Villani. Optimal transport and curvature. In *Nonlinear PDE's and applications*, pages 171–217. Springer, 2011.

[5] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. 2001.

[6] Youssef Mroueh. Wasserstein style transfer. *arXiv preprint arXiv:1905.12828*, 2019.

[7] R Flamary, N Courty, D Tuia, and A Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell*, 2016.

[8] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.