The Wasserstein 1 Distance - Constructing an Optimal Map and Applications to Generative Modelling

Tristan Milne

March 12th

Abstract

Recent advances in generative modelling have shown that machine learning algorithms are capable of generating high resolution images of fully synthetic scenes which some researchers call “dreams” or “hallucinations” of the algorithm. Poetic language aside, one needs to look no further than deepfake videos to know that this technology is shockingly effective and has profound societal implications.

In this talk I will describe the theoretical underpinnings for a highly successful generative modelling technique known as Wasserstein Generative Adversarial Networks (WGANs). These are related to the Wasserstein 1 distance between two probability distributions, which was the original optimal transport problem posed by Monge in 1781 and uses the Euclidean distance between two points to measure the cost of transport. In the first part of my talk I will outline the standard method for constructing an optimal map for this cost, which has beautiful geometric structure. I'll also compare the Wasserstein 1 distance to other Wasserstein distances using powers of the Euclidean cost, explaining why the former is hugely popular in generative modelling while the latter techniques have been mostly ignored in this context.

In the second part of this talk I'll describe the algorithm behind WGANs, show some numerical results, and touch on some open questions related to these techniques.

1 Motivation

Suppose we are given a compact set $\Omega \subset \mathbb{R}^n$ and a target probability distribution $\nu \in \mathcal{P}(\Omega)$. Although ν is *a priori* unknown to us, we will assume that we have a set of samples $\{y_j\}_{j=1}^M$ from ν. However, it may be the case that producing samples from ν is a difficult, expensive, or time-consuming process. For example

1. If ν governs a certain class of images that we want to generate for a computer graphics application, the creation of any one image can require many days and significant expense (c.f. Shrek's Law), or
2. Access to ν may be prevented by a corporation, as in this example, where an anti-money laundering start-up was in need of data normally available only to banks to demo their product.

Suppose further that we have a parametrized probability distribution μ_w which we can sample. For example, we might have a parametrized function $G_w : \mathbb{R}^m \to \mathbb{R}^d$ and feed it noise vectors $z \in \mathbb{R}^m$ distributed according to $\zeta = \mathcal{N}(0, I_m)$, the standard normal distribution on $\mathbb{R}^m$. Then $x = G_w(z)$ is itself a random variable with distribution $\mu = (G_w)_\# \zeta$ which we can sample by sampling z and then computing $G_w(z)$. We then wish to tune w so that $(G_w)_\# \zeta$ and ν are close in some sense; provided our definition of "closeness" is a good one, this will mean that we can sample ν by instead sampling $(G_w)_\# \zeta$.

So how do we measure "closeness" of two probability distributions? There are many possibilities; the first idea in the current wave of interest in generative modelling was to use the Kullback-Liebler divergence [8]. In these notes I'll focus on using the Wasserstein 1 distance and explain a technique known as Wasserstein Generative Adversarial Networks (WGANs) with Gradient Penalty [9].

This technique has been hugely successful in generative modelling. For example, photo realistic images of fake people have been generated using it; you can view some examples here.

2 Structure of these notes

These notes are intended as a companion document to the slides which I presented in my talk, and will follow the same structure. My talk contains no new results, and as such I will only write proofs in cases where they are especially short or illuminating, giving references as appropriate so that the reader may read full proofs elsewhere if they wish.

In Section 3 I will provide some background from Optimal Transport and compare the relative advantages of different costs as they relate to applications. Section 4 contains an explanation of the standard method of how to construct an optimal transport map for the Wasserstein 1 cost using the transport rays of an optimal potential. Finally, in Section 5 I'll explain the algorithm behind WGANs and touch on some open questions and ethical issues.

3 Background on Optimal Transportation

I will assume the reader is somewhat familiar with optimal transportation and as such will be terse here. For an excellent introduction to the subject, see [13]; I really can't recommend this book highly enough, and I have drawn on it extensively for these notes.

Definition 1. *Let $\Omega \subset \mathbb{R}^d$ be a compact set, and let $\mu, \nu \in \mathcal{P}(\Omega)$ be probability measures. Let $c : \Omega \times \Omega \to \mathbb{R}$ be a cost function, and define the Monge Problem (MP) as*

$$\inf_{T_{\#}\mu=\nu} \int_\Omega c(x, T(x)) d\mu,$$

where $T_{\#}\mu$ denotes the pushforward measure defined by

$$T_{\#}\mu(E) = \mu(T^{-1}(E)).$$

A map T realizing the infimum in (MP) is called an optimal map.

Definition 2. *The Kantorovich Problem (KP) is*

$$\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\Omega^2} c(x, y) d\gamma, \quad \Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(\Omega \times \Omega) \mid (\pi_x)_{\#}\gamma = \mu, (\pi_y)_{\#}\gamma = \nu\}.$$

A $\gamma \in \Pi(\mu, \nu)$ realizing the infimum in (KP) is called an optimal plan.

If one thinks of γ as a multi-valued map such that $\gamma(E_1 \times E_2)$ is the amount of mass that γ takes from E_1 and moves to E_2, the constraint that $\gamma \in \Pi(\mu, \nu)$ just means that the mass that γ takes from a set $E \subset \Omega$ is the amount of mass that μ puts there, while the mass that γ gives to a set E is the amount of mass that ν puts there.

A few remarks are in order on how these problems relate to each other.

- While the admissible set in (MP) is possibly empty (there is no map which pushes δ_0 onto $(1/2)\delta_0 + (1/2)\delta_1$), the admissible set in (KP) is always non-empty because it contains the product measure $\mu \otimes \nu$.

- On the other hand, if T is an admissible map for (MP), then the measure $(I,T)_{\#}\mu$ is admissible for (KP) and obtains the same value. This implies that the infimum of (MP) is no less than that of (KP).

The most useful thing about (KP) is that it is a linear optimization problem on the space $\mathcal{P}(\Omega \times \Omega)$. This allows for straightforward proofs of the existence of an optimal plan and a duality result which is essential to the theory.

Theorem 1. *If $c : \Omega \times \Omega \to \mathbb{R}$ is continuous, then (KP) admits a solution (i.e. an optimal plan exists).*

Proof. See Theorem 1.4 in [13]. □

There are more general existence theorems with weaker assumptions on the cost, but they will not be necessary for our purposes.

One facet of optimal transport research in the late 20th century was the problem of showing, for various cost functions, that one could obtain an optimal plan which is induced by a map, i.e. $\gamma = (I,T)_{\#}\mu$. If it exists, this map is automatically optimal for (MP), and a proof of the existence of such a map would therefore close the loop initially opened by Kantorovich. It turns out that the following duality result is essential for closing this loop.

Theorem 2. *If Ω is compact and $c : \Omega \times \Omega \to \mathbb{R}$ is continuous, then*

$$\min_{\gamma \in \Pi(\mu,\nu)} \int_{\Omega^2} c(x,y)d\gamma = \max_{\varphi,\psi \in C(\Omega), \varphi \oplus \psi \leq c} \int_\Omega \varphi d\mu + \int_\Omega \psi d\nu.$$

Here, $\varphi \oplus \psi$ is the function on $\Omega \times \Omega$ given by

$$\varphi \oplus \psi(x,y) = \varphi(x) + \psi(y).$$

A pair of functions (φ, ψ) solving the dual problem are known as Kantorovich potentials, or just potentials for short.

Proof. To show this formally, do what is usually done in linear programming and write (KP) as

$$\inf_{\gamma \in \mathcal{M}(\Omega \times \Omega)} \int_{\Omega^2} c(x,y)d\gamma + \sup_{\varphi,\psi \in C(\Omega)} \int_\Omega \varphi(x)(d\mu - d(\pi_x)_{\#}\gamma) + \int_\Omega \psi(y)(d\nu - d(\pi_y)_{\#}\gamma)$$

Formally interchanging the infimum and the supremum and taking the infimum over γ produces the dual problem. For a rigorous (and slick!) proof in the compact case, see Section 1.6.3 of [13]. □

Via the c-transform we can halve the number of variables in the dual problem and eliminate the constraint. To simplify notation, we will assume from this point onward that $c(x, y)$ is symmetric in x and y (i.e. $c(x, y) = c(y, x)$).

Proposition 1. *For $c : \Omega \times \Omega \to \mathbb{R}$ a continuous cost, define the c-transform of a function $\varphi : \Omega \to \mathbb{R}$ as $\varphi^c : \Omega \to \mathbb{R}$,*

$$\varphi^c(y) = \inf_{x\in\Omega} c(x, y) - \varphi(x).$$

Then

$$\max_{\varphi,\psi\in C(\Omega),\varphi\oplus\psi\leq c} \int_\Omega \varphi d\mu + \int_\Omega \psi d\nu = \max_{\varphi\in C(\Omega)} \int_\Omega \varphi d\mu + \int_\Omega \varphi^c d\nu$$

Proof. The constraint on the pair (φ, ψ) is

$$\varphi(x) + \psi(y) \leq c(x, y), \quad \forall x, y \in \Omega.$$

In order for (φ, ψ) to solve the given maximization problem, it's clear that given φ, ψ should saturate this constraint as long as continuity is retained. Hence, given y we may as well take

$$\psi(y) = \varphi^c(y) = \inf_{x\in\Omega} c(x, y) - \varphi(x).$$

The function φ^c can be shown to be continuous by showing that it inherits the continuity of c; see Box 1.8 in [13] for a rigorous proof. □

We need one final result before we can state how potentials can be used to find an optimal map.

Definition 3. *We say that $\varphi \in C(\Omega)$ is c-concave if there exists $\psi : \Omega \to \bar{\mathbb{R}}$ such that*

$$\varphi = \psi^c.$$

We denote the set of all c-concave functions on Ω as c-conc(Ω).

Proposition 2. *For $\varphi \in C(\Omega)$,*

$$\varphi^{cc} \geq \varphi, \quad \varphi^{ccc} = \varphi^c. \tag{1}$$

As such

$$\max_{\varphi,\psi\in C(\Omega),\varphi\oplus\psi\leq c} \int_\Omega \varphi d\mu + \int_\Omega \psi d\nu = \max_{\varphi\in c\text{-}conc(\Omega)} \int_\Omega \varphi d\mu + \int_\Omega \varphi^c d\nu \tag{2}$$

Proof. For a proof of (1), see Proposition 1.34 from [13]. To see that this implies (2), we observe that given a pair (φ, φ^c), $(\varphi^{cc}, \varphi^{ccc})$ is still admissible and (1) implies that it obtains a value of the dual functional no less than the original pair. Making our new variable φ^{cc}, we have the result. □

Since we have reduced the dual problem down to one variable φ, we will call φ a Kantorovich potential, or just a potential for short, if (φ, φ^c) solves the dual problem. We will also often take φ c-concave implicitly.

Duality gives the following result which will be used to find an optimal map given a potential.

Proposition 3. *Suppose that φ is a Kantorovich potential and γ is an optimal plan. Then*

$$spt(\gamma) \subset \{(x, y) \in \Omega \times \Omega \mid \varphi(x) + \varphi^c(y) = c(x, y)\}$$

Proof. By definition of the c-transform, we have

$$\varphi(x) + \varphi^c(y) \leq c(x, y). \tag{3}$$

Integrating both sides of this equation with respect to γ, we obtain

$$\int_\Omega \varphi(x)d\mu + \int_\Omega \varphi^c(y)d\nu \leq \int_{\Omega^2} c(x, y)d\gamma.$$

But since γ is an optimal plan and φ is a potential, Theorem 2 (the duality result) implies that this inequality must be equality. hence (3) must be an equality γ almost everywhere, and continuity of the functions involved implies that this equality extends to the support of γ. □

Proposition (3) gives a hint of how to construct a map given a potential; if $(x, y) \in \text{spt}(\gamma)$ for γ an optimal plan then,

$$\varphi(x) + \varphi^c(y) = c(x, y).$$

Rearranging and recalling the definition of the c-transform, we get

$$\varphi^c(y) = c(x, y) - \varphi(x) = \min_{z\in\Omega} c(z, y) - \varphi(z).$$

Hence, x obtains the minimizer in the c-transform $\varphi^c(y)$. If the set of y for which x is in this argmin is unique, then γ is single valued at x (i.e. it is induced by a map there). Thus, to proceed with the construction of an optimal map we will need additional assumptions on c which guarantee this property.

3.1 The case $c(x, y) = |x - y|^p$

The case of $c(x, y) = |x - y|^p$, where $|\cdot|$ denotes the Euclidean norm and $p \geq 1$ is frequently considered. In this case, (KP) becomes

$$\min_{\gamma\in\Pi(\mu,\nu)} \int_{\Omega^2} |x - y|^p d\gamma.$$

We use the notation $W_p(\mu, \nu)$ to denote the pth root of this quantity, and call W_p the Wasserstein p distance; it is a fact that W_p is a metric on $\mathcal{P}(\Omega)$ (Proposition 5.1, [13]).

There are physical interpretations for this quantity for different values of p.

- If $p = 1$, $\int_{\Omega^2} |x - y| d\gamma$ is a measure of the distance travelled by each particle of mass times the mass of that particle. Since the force required to lift the particle is proportional to its mass, this quantity measures the work required to move μ onto ν with plan γ. This was the first case considered by Monge, who might have been concerned with how many peasants would be needed to dig a trench.
- If $p = 2$, $\int_{\Omega^2} |x - y|^2 d\gamma$ can be interpreted as proportional to the kinetic energy of a mass of fluid initially distributed according to μ where each particle moves with velocity $|x - y|$.

3.2 Constructing an optimal map for $p > 1$

It turned out that constructing an optimal map was easier for the case $p > 1$ due to the strict convexity of the cost; the method I'll sketch here is due to Brenier ([2] and [3]), later generalized by Gangbo and McCann ([6]) to the case of $c(x, y) = h(x - y)$ for h strictly convex.

Theorem 3. *If $\mu \ll \mathcal{L}$ (Lebesgue measure on $\mathbb{R}^d$) and $\partial\Omega$ is negligible, then there is a unique optimal plan γ for (KP) with $p > 1$. It is of the form $\gamma = (I, T)_{\#}\mu$, where T satisfies*

$$T(x) = x - (\nabla | \cdot |^p)^{-1}(\nabla \varphi(x)),$$

where φ is a potential.

Proof. As pointed out above, for all $(x, y) \in \mathrm{spt}(\gamma)$,

$$x \in \mathrm{argmin}_z |z - y|^p - \varphi(z)$$

As such, if x is a point where φ is differentiable[1] then

$$\nabla(| \cdot - y|^p)(x) = \nabla \varphi(x).$$

[1] Guaranteeing that φ is differentiable μ almost everywhere is where the assumptions on μ and $\partial\Omega$ come in.

Since $p > 1$, $x \mapsto |x|^p$ is a strictly convex function and hence its gradient vector field is injective. As such,

$$x - y = (\nabla|\cdot|^p)^{-1}(\nabla\varphi(x)),$$

and rearranging gives y uniquely as a function of x. For a rigorous proof, see Theorem 1.17 from [13]. □

Because the gradient field of $x \mapsto |x|$ is not injective, the same proof is not possible for the case of $p = 1$. In fact, there is no simple formula relating an optimal potential to an optimal map for the case $p = 1$. But the case $p > 1$ doesn't have it so easy, since we have offloaded the difficulty of computing an optimal transport map to the computation of an optimal potential. As the next section will show, it is much more difficult to compute an optimal potential for the case $p > 1$ than it is for the case $p = 1$.

3.3 Computing the c-transform for $p = 1$ vs $p > 1$

Computing an optimal potential requires for us to solve

$$\max_{\varphi\in c\text{-conc}(\Omega)} \int_\Omega \varphi d\mu + \int_\Omega \varphi^c d\nu$$

for φ. If one wishes to do so using, say, a gradient scheme in φ, it would be necessary to compute φ^c. It turns out that this is spectacularly easy for $p = 1$, and much more difficult for $p > 1$. I'll treat the case of $p = 1$ first.

Proposition 4. *For $p = 1$, φ^c is 1-Lipschitz, denoted $\varphi \in$ 1-Lip(Ω). Moreover, if $\varphi \in$ 1-Lip(Ω), then $\varphi^c = -\varphi$. Hence*

$$c\text{-}conc(\Omega) = 1\text{-}Lip(\Omega).$$

Proof. That $\varphi^c \in$ 1-Lip(Ω) follows because φ^c inherits the modulus of continuity of $c(x, y) = |x - y|$ (see Box 1.8 from [13]). This proves one inclusion of (4). For the second statement, let $\varphi \in$ 1-Lip(Ω) and compute

$$\varphi^c(y) = \min_z |z - y| - \varphi(z) \geq -\varphi(y),$$

with the minimum clearly obtained at $z = y$. Thus, $\varphi = -(-\varphi)$ implies that every 1 Lipschitz function is c-concave. □

Hence, if φ is already 1-Lipschitz, then φ^c is computed with no effort. This is not the case with $p > 1$; I'll demonstrate this for the case $p = 2$, but similar computational complexity arguments for different costs (albeit ones that separate along coordinates) can be found in [10], Section 4.1.

Proposition 5. *If $p = 2$, computing φ^c is equivalent to computing a Legendre transform. On a discrete grid in $\mathbb{R}^d$ with n points per dimension, this can be done in $O(n^d)$ operations.*

Proof. We compute

$$\begin{aligned}\varphi^c(y) &= \min_{x\in\Omega} |x-y|^2 - \varphi(x),\\ &= |y|^2 - \max_{x\in\Omega}\langle 2y, x\rangle - (|x|^2 - \varphi(x)).\end{aligned}$$

The maximization problem is the Legendre transform of the function $x \mapsto |x|^2 - \varphi(x) + 1_\Omega(x)$, which can be computed on the grid in question in $O(n^d)$ computations [12]. □

If d is large, as it is in the case of image generation where d is three times the number of pixels in an image, then computing any single Legendre transform at this complexity is totally infeasible. As we will see later, this has motivated interest for the machine learning community in the W_1 distance.

4 Constructing an Optimal Map for $W_1(\mu, \nu)$

In this section I'll sketch a method for proving the existence of an optimal map for $W_1(\mu, \nu)$.

4.1 A brief history of solutions

The first partial proof of existence came from [14], however it was later discovered to have a gap which was eventually filled by L. Ambrosio. In the intervening time other authors tackled the problem, and the first complete proof appears to have been from Evans and Gangbo [5]. Their method of proof computes an optimal transport map as the flow a certain vector field related to an optimal potential; however, to make their results work they needed that both μ and ν be absolutely continuous with respect to Lesbesgue measure and have Lipschitz densities. Soon after came results from Caffarelli, Feldman, and McCann [4] and Trudinger and Wang [15], both of which apply to the case of μ and ν absolutely continuous but without the additional Lipschitz assumption. In addition, [4] treats the case of $c(x, y) = \|x - y\|$ a general norm with a strictly convex and smooth unit ball (i.e. not necessarily Euclidean). Additionally, a proof for a more general case (with only $\mu \ll \mathcal{L}$) is given in [13], Section 3.1 with a method due to L. Ambrosio.

Though I will state the result in the case of only $\mu \ll \mathcal{L}$, the argument I'll sketch is that of [4] and [15], which I think captures the essential ideas more clearly.

Theorem 4. *If $\mu \ll \mathcal{L}$, there is an optimal transport map T for $W_1(\mu,\nu)$.*

The strategy is to use Proposition 3 to determine the support of an optimal plan γ. This will decompose the domain into essentially disjoint segments known as "transport rays", and after solving a one dimensional transport problem on each of the segments we will have an optimal map.

4.2 Transport rays

When working with the $p = 1$ case we will use the notation u for a potential, as opposed to φ. We begin by translating Proposition 3 to our setting.

Proposition 6. *Suppose that u is a Kantorovich potential and γ is an optimal plan for $W_1(\mu,\nu)$. Then*

$$\mathit{spt}(\gamma) \subset \{(x,y) \in \Omega \times \Omega \mid u(x) - u(y) = |x-y|\}$$

Let us examine the set $T = \{(x,y) \in \Omega \times \Omega \mid u(x) - u(y) = |x-y|\}$ further.

Proposition 7. *If $u \in 1\text{-}\mathit{Lip}(\Omega)$ and*

$$u(x) - u(y) = |x-y|,$$

then for all $z \in [x,y] := \{(1-t)x + ty \mid t \in [0,1]\}$,

$$u(x) - u(z) = |x-z|. \tag{4}$$

Proof. We have

$$\begin{aligned}
|x-y| &= u(x) - u(y), \\
&= u(x) - u(z) + u(z) - u(y), \\
&\leq |x-z| + |z-y|, \\
&= |x-y|
\end{aligned}$$

So every inequality is equality, giving (4). □

Thus, if $(x,y) \in T$, then for every pair of points $w, z \in [x,y]$ with $u(w) \geq u(z)$ we also have $(w,z) \in T$. We give these segments a special name.

Definition 4. *We call a non-singleton segment $[x, y]$ a transport ray of u if*

$$u(x) - u(y) = |x - y|$$

and $[x, y]$ is the largest such segment containing x and y.

Example 1. *Let Ω be the unit ball centred at the origin, and take $u(x) = dist(x, \partial B_1(0))$, where dist denotes the Euclidean distance. This function is 1-Lipschitz, and the transport rays of u are all segments of the form $[0, x]$ where $x \in \partial B_1(0)$.*

On the interior of a ray we have differentiability of u. Moreover, the gradient of u points in the direction of the ray, which allows us to show that the rays are essentially disjoint from each other.

Proposition 8. *Let $[x, y]$ be a transport ray of a 1-Lipschitz function u. Then for all z on the interior of the ray (i.e. $z \in \{(1-t)x + ty \mid t \in (0, 1)\} =:]x, y[$), $\nabla u(z)$ exists and satisfies*

$$\nabla u(z) = \frac{x - y}{|x - y|}.$$

As such, two transport rays can only intersect at their endpoints, and the Lebesgue measure of the set of points in more than one ray is 0.

Proof. The differentiability of u on ray interiors and the value of the gradient is shown in Lemma 3.6 of [13]. Using this I'll show the second part of the proposition.

Suppose that two rays $[x_1, y_1]$ and $[x_2, y_2]$ intersect at a point $z \in]x_1, y_1[$. Then u is differentiable at z with

$$\nabla u(z) = \frac{x_1 - y_1}{|x_1 - y_1|}.$$

But since $z \in [x_2, y_2]$ we can take a directional derivative in the direction $(x_2 - y_2)/|x_2 - y_2|$ and use the norm of $\nabla u(z)$ to obtain that

$$\nabla u(z) = \frac{x_2 - y_2}{|x_2 - y_2|}.$$

Thus, the directions of the rays $[x_1, y_1]$ and $[x_2, y_2]$ are the same, and they share a point z. By maximality of transport rays we must have that they are the same.

If z is in more than one transport ray, then u cannot be differentiable at z. As such, Rademacher's theorem gives that the set of points in more than one transport ray has Lebesgue measure 0. □

The preceding proposition allows us to decompose Ω into the set of transport rays, which are disjoint up to a negligible set. I am leaving out some details here (for example, how do I know that every point is inside some transport ray? Answer: it's not), but I welcome the reader to consult the reference [4] for an explanation.

4.3 Construction of an optimal map

I'll start by observing that if T is an admissible map that preserves transport rays, it must be optimal.

Proposition 9. *If T satisfies $T_{\#}\mu = \nu$, u is a potential for $W_1(\mu, \nu)$, and*

$$u(x) - u(T(x)) = |x - T(x)| \tag{5}$$

μ almost everywhere, then T is an optimal map.

Proof. Integrate both sides of (5) with respect to μ to get

$$\int_{\Omega} u(x)d\mu - \int_{\Omega} u(T(x))d\mu = \int_{\Omega} |x - T(x)|d\mu$$

Changing variables in the second integral and using $T_{\#}\mu = \nu$, this becomes

$$\int_{\Omega} u(x)d\mu - \int_{\Omega} u(y)d\nu = \int_{\Omega} |x - T(x)|d\mu$$

The left hand side is $W_1(\mu, \nu)$ by assumption on u, and this shows that T is optimal. □

The preceding proposition gives us a strategy for constructing an optimal map; we need to find a map which preserves transport rays as well as balances mass (i.e. $T_{\#}\mu = \nu$). Given a map T which preserves transport rays, it's possible to show that if T satisfies an appropriate mass balance on each ray it will balance mass globally. I will not state this result formally in these notes; consult [4] for a full statement.

The advantage of restricting to the rays is that the mass balance condition $T_{\#}\mu = \nu$, which is quite intractable for high dimensions, becomes easy in one dimension.

Theorem 5. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$ and define the cumulative distribution functions*

$$F_{\mu}(x) = \mu((-\infty, x]), \quad F_{\nu}(y) = \nu((-\infty, y])$$

If μ is atomless (i.e. it assigns no measure to any single point), the map

$$T(x) = \inf\{t \in \mathbb{R} \mid F_{\nu}(t) \geq F_{\mu}(x)\}$$

satisfies $T_{\#}\mu = \nu$.

The preceding theorem gives us a way to determine the map on each ray. Suppose now we were in in the simple case where every transport ray is vertical, (i.e. $u(x_1, \ldots, x_n) = x_n$). Then we could obtain an optimal map by fixing the first $n-1$ variables and sending x_n to the value stipulated by Theorem 5 so that mass is balanced on that ray.

In the case of rays which are not straight, the authors of [4] applied a local change of coordinates to straighten out the rays; the point corresponding to x in the new coordinates has last coordinate specified by $u(x)$, with the remaining coordinates parametrizing the level sets of u. It is precisely the decomposition of the domain into transport rays that makes such a change of variable possible, but additional regularity results on the twisting of rays are necessary to make the argument rigorous; these are provided in Lemma 16 of [4].

4.4 Some loose ends

In the preceding section I have laid out a sketch of how to construct an optimal map for $W_1(\mu, \nu)$. There are several details which I have glossed over, including

- How do we determine the map T on points which are not in any transport ray?
- What exactly is meant by balancing mass on each ray?
- Is it true that balancing mass on each ray will lead to $T_{\#}\mu = \nu$?
- How is the the local change of variables that straightens out the rays constructed?
- Is the map T we have constructed by considering each ray individually a measurable map?

If you are interested in the answers to these questions, I encourage you to read [4] in detail.

4.5 A final observation

Now that we have the existence of an optimal map, we can make an additional observation. Recall that we have previously stated that there is no simple formula for the optimal transport map T for the case of $p = 1$ in terms of the potential u alone. That does not mean that we cannot glean any information regarding T from u directly. Indeed, if x is a point where u is differentiable, then since $(x, T(x))$ is on a transport ray,

$$\nabla u = \frac{x - T(x)}{|x - T(x)|}.$$

Thus, ∇u gives us the direction of transport, but not the distance.

5 Wasserstein GANs

In this section I will explain the algorithm for WGANs with Gradient Penalty (WGAN-GP), first appearing in [9].

5.1 Recalling the problem

Recall that our intention was to measure the closeness of $\mu := (G_w)_{\#}\zeta$ and ν, with the intent of using this metric as a way of finding good parameters w so that $\mu \approx \nu$. The authors of [9] use the W_1 distance as their metric; as we will shortly see, the simplicity of the c transform in this case explains this choice of metric as opposed to W_p for $p > 1$.

We must now answer several questions

1. How do we design G_w so as to have a hope of approximating ν? Surely if G_w is too simple (i.e. a linear function), no choice of w will allow us to approximate ν.
2. Given G_w, how do we compute $W_1((G_w)_{\#}\mu, \nu)$?
3. Given $W_1((G_w)_{\#}\mu, \nu)$, how do we tune w to as to move reduce this distance?

I will answer each of these in order.

5.2 Designing G_w

The function $G_w : \mathbb{R}^m \to \mathbb{R}^d$, called the generator, will be parametrized by a feedforward neural network. Rather than attempt to give a comprehensive definition of neural networks and explain the nuances of their design, I will give a rough description of what they are which should be satisfying enough for these notes. If the reader wants more information, they should consult Chapter 6 of [7].

- Feedforward neural networks are constructed by composing several simple functions (called "layers of the network") which are usually made up of an affine map followed by a non-linear function. A popular choice for this simple function is

  $$f(x) = \sigma(Wx + b), \quad \sigma(z_1, \ldots, z_n) = (z_1^+, \ldots, z_n^+), \quad z_i^+ = \max(0, z_i).$$

  where (W, b) are parameters specifying the affine map. The parameters for all the layers will make up the full parameter vector w.

- Given enough parameters, it is possible to show that a neural network can approximate any continuous function ([11]).
- The process of determining a set of parameters w so that G_w performs well at a given task is known as "training the network". This is often accomplished by applying stochastic gradient descent to a loss function which measures performance on the task, and there are several convenient software packages for doing this (e.g. PyTorch, TensorFlow).
- A special type of feedforward neural network known as a convolutional neural network (CNN) excels at tasks relating to analysis of images. For a CNN, the general linear maps W are replaced by matrices associated with discrete convolutions.

5.3 Computing $W_1((G_w)_{\#}\mu, \nu)$

Given G_w, we will compute $W_1((G_w)_{\#}\mu, \nu)$ by the dual formulation

$$W_1(\mu, \nu) = \sup_{u \in 1\text{-Lip}(\Omega)} \int_\Omega u(d\mu - d\nu)$$

We will attempt to solve the dual problem by parametrizing u by yet another neural network u_θ,

$$W_1((G_w)_{\#}\zeta, \nu) = \sup_{\theta, u_\theta \in 1-\text{Lip}(\Omega)} \int_\Omega u_\theta(d(G_w)_{\#}\zeta - d\nu).$$

The question of which parameters θ lead to a 1-Lipschitz u_θ is non-trivial given the complicated parametrization of u_θ. The initial idea in the seminal WGAN paper [1] was to bound the norms of the affine maps in the definition of u_θ, however this was found to produce an insufficiently rich class of functions. The method we are discussing here is an improvement on [1] which adds a regularization to penalize large gradients of u. The ideal loss function for u_θ - the function which we want to minimize over the parameters θ - in [9] is

$$L(\theta) = \int_\Omega u_\theta(d\nu - d(G_w)_{\#}\zeta) + \lambda R[\nabla u_\theta], \quad R[\nabla u_\theta] = \int_\Omega (\|\nabla u_\theta(z)\| - 1)^2 d\sigma(z),$$

where σ is a distribution that I will not specify. The particular regularization R is selected because, as we know from Proposition 8, the norm of ∇u is equal to 1 on the interior of transport rays, and this regularizer penalizes any gradient which does not have norm 1.

I use the phrase "ideal" loss because in practice we cannot compute the integrals in L, we can only approximate them with samples.

With that, we are ready to state the algorithm used to compute u_θ and thus $W_1((G_w)_\#\zeta, \nu)$. In practice, a more complicated version of gradient descent known as Adam is used in step 5, however an explanation of that algorithm is beyond the scope of these notes.

Algorithm 1: Given a current value of $w = w_0$ and $\theta = \theta_0$, batch size N and step size η,

1. Generate fake samples $\{x_i\}_{i=1}^N$, $x_i = G_{w_0}(z_i)$, $z_i \sim \mathcal{N}(0, I_m)$.
2. Take real samples $\{y_i\}_{i=1}^N$ from ν.
3. For each pair x_i, y_i, sample $t_i \sim U([0,1])$
4. Approximate L with samples

$$\begin{aligned} L(\theta_0) &\approx \frac{1}{N}\sum_{i=1}^N u_{\theta_0}(y_i) - u_{\theta_0}(x_i) + \lambda(||\nabla u_{\theta_0}((1-t_i)x_i + t_i y_i)|| - 1)^2, \\ &=: \hat{L}(\theta_0) \end{aligned}$$

5. Update θ_0 by gradient descent

$$\theta_0^{\text{new}} = \theta_0 - \eta \nabla \hat{L}(\theta_0).$$

6. Repeat steps 1 - 5 until the value of $\hat{L}(\theta_0)$ stabilizes or until a pre-set maximum number of iterations.

The sampling used for the gradient penalty term in $\hat{L}$ is used to probe the convex hull of the supports of μ and ν, which contains the region over which transport occurs.

5.4 Tuning w to reduce $W_1((G_w)_\#\zeta, \nu)$

The parameters of the generator G_w are trained by using the estimated value of $W_1((G_w)_\#\zeta, \nu)$ as a measure of the performance of w; by decreasing this value, we move $(G_w)_\#\zeta$ closer to ν. Just like the last section, this is accomplished by gradient descent (or a variant like Adam) on estimates of $W_1((G_w)_\#\zeta, \nu)$ using samples.

Algorithm 2: Given initial parameters w_0 and gradient descent step size ϵ,

1. Compute u_θ using Algorithm 1,

2. Estimate $W_1((G_{w_0})_\# \zeta, \nu)$ using u_θ and samples $\{G_{w_0}(z_i)\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$,

$$W_1((G_{w_0})_\# \zeta, \nu) \approx \frac{1}{N} \sum_{i=1}^N u_\theta(G_{w_0}(z_i)) - u_\theta(y_i) =: \hat{L}'(w_0)$$

3. Perform gradient descent on $\hat{L}'(w_0)$

$$w_0^{\text{new}} = w_0 - \epsilon \nabla_w \hat{L}'(w_0) \tag{6}$$

4. Repeat steps 1-3 until samples $G_{w_0}(z)$ are of sufficient visual quality.

5.5 Open questions

While reading the preceding sections you may have found yourself asking many questions about this technique; this field is so new that likely most of your questions do not have satisfying answers beyond “empirical evidence shows that this is a good technique”, if you can even call that satisfying. Here are some open questions that I think about in my work.

- The optimization problems for finding w and θ are massively high dimensional and non-convex; why does gradient descent with sampling (also known as stochastic gradient descent) work so well?
- Does solving
$$\min_\theta \int_\Omega u_\theta(d\nu - d(G_{w_0})_\# \zeta) + \lambda R[\nabla u_\theta]$$
actually produce a Kantorovich potential?
- The W_1 distance is known to have sample complexity - a measure of how well $W_1(\mu, \nu)$ can be approximated by samples - that scales exponentially with dimension [16]; how do we get good results despite this?
- In reality we do not train θ to completion before updating w; how do the dynamics of these two descent schemes affect each other?

We should also consider the ethical implications of this technology. If realistic synthetic data can be easily generated, one could imagine bad actors using this technology to, for example, generate images of individuals doing or saying things that they have not done in reality; in fact, this is already being done. As such, it is an ethical imperative for researchers in this area to consider how their work might be misused and possibly develop safeguards to prevent this.

6 Brief Summary of Talk

In these notes I have discussed the W_1 distance and an application of it to generative modelling.

- After recalling some background in optimal transport, we discussed how W_p for $p > 1$ compares to W_1
  - With $p > 1$, a potential gives an optimal map, whereas for $p = 1$ a potential gives only the direction of transport.
  - With $p = 1$, the c-transform is easier to compute; this is one reason why $p = 1$ is more popular in generative modelling.
- We sketched a method for constructing an optimal map for $W_1(\mu, \nu)$
  - We decompose the space into transport rays, and solve the resulting 1-D problems.
- We went over the algorithm for training WGANs and touched on some open questions and ethical concerns.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[2] Y. Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305:805–808, 1987.

[3] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[4] L. Caffarelli, M. Feldman, and R. McCann. Constructing optimal maps for monge's transport problem as a limit of strictly convex costs. *Journal of the American Mathematical Society*, 15(1):1–26, 2002.

[5] L. C. Evans and W. Gangbo. *Differential equations methods for the Monge-Kantorovich mass transfer problem.* Number 653. American Mathematical Soc., 1999.

[6] W. Gangbo and R. J. McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.

[7] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[9] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[10] M. Jacobs and F. Léger. A fast approach to optimal transport: The back-and-forth method. *Numerische Mathematik*, 146(3):513–544, 2020.

[11] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 1993.

[12] Y. Lucet. Faster than the fast legendre transform, the linear-time legendre transform. *Numerical Algorithms*, 16(2):171–185, 1997.

[13] F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55:58–63, 2015.

[14] V. N. Sudakov. *Geometric problems in the theory of infinite-dimensional probability distributions*, volume 141. American Mathematical Soc., 1979.

[15] N. S. Trudinger and X.-J. Wang. On the monge mass transfer problem. *Calculus of Variations and Partial Differential Equations*, 13(1):19–31, 2001.

[16] J. Weed, F. Bach, et al. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.