A Brief Introduction to Optimal Transport Theory

D. P. Bourne*

July 27, 2018

Abstract

These lecture notes are for a short course given at the London Mathematical Society Undergraduate Summer School at the University of Glasgow, 26-27 July 2018.

1 Motivation

The aim of these notes is to give penultimate year undergraduate students a flavour of the fashionable research area of optimal transport theory. The origin of the subject goes back to 1781 and the French engineer Gaspard Monge [7], who was interested in the problem of the optimal way of redistributing mass, e.g., given a pile of soil, how can it be transported and reshaped to form an embankment with minimal effort? This problem remained unsolved for over 200 years (it was not even known whether *there existed* an optimal way of redistributing mass), until some big mathematical breakthroughs in the 1980s and 1990s. Since then the field has flourished and optimal transport theory has found applications in PDEs, geometry, statistics, economics and image processing. There are now several excellent textbooks on optimal transport theory for PhD students and researchers [1, 2, 5, 8, 9, 10, 11, 12, 13]. As far as I am aware, however, there are currently no textbooks targeted at undergraduate students, and optimal transport is not typically taught at the undergraduate level in the UK. In these lecture notes I aim to give an accessible introduction to the subject without assuming any background knowledge in measure theory, which is usually a prerequisite.

2 Notation and Background Material

Throughout this course we will use the following notation:

- **Characteristic functions.** Let $A \subset \mathbb{R}^d$. The *characteristic function* $\chi_A : \mathbb{R}^d \to \{0,1\}$ is

$$\chi_A(x) := \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

- **Preimage.** Let $T : X \to Y$, $B \subseteq Y$. The *preimage* of B under T is the set

$$T^{-1}(B) := \{x \in X : T(x) \in B\}.$$

*Durham University

- **Probability densities.** Let $X \subseteq \mathbb{R}^d$. We say that $f : X \to [0, \infty)$ is a *probability density* on X if $\int_X f(x)\,\mathrm{d}x = 1$.
- **Push-forward.** Let $X, Y \subseteq \mathbb{R}^d$ and $T : X \to Y$. Let f be a probability density on X and g be a probability density on Y. We say that g is the *push-forward* of f under T, and write $g = T\#f$, if

$$\int_B g(y)\,\mathrm{d}y = \int_{T^{-1}(B)} f(x)\,\mathrm{d}x \quad \forall\, B \subseteq Y. \tag{2.1}$$

  In other words, the mass of the set B with respect to the density g equals the mass of the set $T^{-1}(B)$ with respect to the density f.

Exercise 2.1. Let $T : \mathbb{R} \to \mathbb{R}$ be the translation $T(x) = x + 1$. Let $f = \chi_{[0,1]}$ and $g = \chi_{[1,2]}$ be probability densities on $\mathbb{R}$. Show that $T\#f = g$. Define $S : \mathbb{R} \to \mathbb{R}$ by $S(x) = 2x$. Show that $S\#f \neq g$.

The following classes of functions play a very important role in optimisation problems:

Definition 2.2 (Convex and concave functions)**.** Let $I \subseteq \mathbb{R}$ be an interval (possibly unbounded). We say that $h : I \to \mathbb{R}$ is *convex* if for all $\lambda \in (0, 1)$, $x, y \in I$, $x \neq y$,

$$h((1 - \lambda)x + \lambda y) \leq (1 - \lambda)h(x) + \lambda h(y). \tag{2.2}$$

We say that h is *strictly convex* if the inequality in (2.2) is strict. We say that h is *concave* if $-h$ is convex and *strictly concave* if $-h$ is strictly convex.

Equation (2.2), convexity of h, means that the graph of h on the interval (x, y) lies below the line joining $h(x)$ to $h(y)$ for all $x, y \in I$. Strict convexity means the graph lies strictly below the line. For concavity and strict concavity, the same is true with *below* replaced by *above*.

Example 2.3. Here are some examples of convex and concave functions on $\mathbb{R}$. The function $h_1(x) = x^2$ is strictly convex, $h_2(x) = |x|$ is convex but not strictly convex, $h_3(x) = ax + b$, $a, b \in \mathbb{R}$, is both convex and concave (but not strictly convex or strictly concave), $h_4(x) = -x^2$ is strictly concave, $h_5(x) = x^3$ is neither convex nor concave.

The following result is useful for checking the convexity of a function:

Theorem 2.4 (Second-derivative test)**.** *Let $h : I \to \mathbb{R}$ be twice differentiable. Then h is convex if and only if $h''(x) \geq 0\ \forall\, x \in I$. If $h''(x) > 0\ \forall\, x \in I$, then h is strictly convex.*

Exercise 2.5 (Strict convexity of h does not imply $h'' > 0$)**.** Find an example of a strictly convex function $h : \mathbb{R} \to \mathbb{R}$ such that $h''(x) = 0$ for some $x \in \mathbb{R}$.

Exercise 2.6. Show that $h_6(x) = x \log x$, $x \in (0, \infty)$, is strictly convex. Show that $h_7(x) = x^{1/2}$, $x \in (0, \infty)$, is strictly concave.

We will also need the following important inequality:

Theorem 2.7 (Jensen's inequality)**.** *Let $h : I \to \mathbb{R}$ be convex, let f be a probability density on $[a, b]$, and let $u : [a, b] \to I$ be bounded. Then*

$$h\left(\int_a^b u(x)f(x)\,\mathrm{d}x\right) \leq \int_a^b h(u(x))f(x)\,\mathrm{d}x.$$

3 The Monge Problem

We are now in a position to state Monge's optimal transport problem in modern mathematical language:

Definition 3.1 (The Monge problem)**.** Let $X, Y \subseteq \mathbb{R}^d$. Let f be a probability density on X and g be a probability density on Y. Let $c : X \times Y \to [0, \infty)$ be continuous. The *Monge problem* is to find a transport map $T : X \to Y$ satisfying $T\#f = g$ such that T minimises the cost functional

$$M(T) := \int_X c(x, T(x)) f(x) \, \mathrm{d}x.$$

The *optimal transport cost* $\mathcal{T}_c(f, g)$ of transporting f to g with cost function c is defined by

$$\mathcal{T}_c(f, g) := \inf_{T\#f=g} M(T).$$

We write *inf* in the definition of $\mathcal{T}_c(f, g)$ rather than *min* since the minimum may not exist (see Example 3.11). In this course we will consider the following fundamental questions: Does there exist an optimal transport map T? If so, is it unique? Can we find an explicit expression for T? If not, can we say something about the properties of T? The answers to these questions will depend on the cost c and the probability densities f and g.

Remark 3.2 (Physical interpretation)**.** Let's interpret Definition 3.1 in terms of Gaspard Monge's original problem of redistributing (transporting and reshaping) a pile of sand or soil to form an embankment with minimal effort: $X = Y = \mathbb{R}^3$; $c(x, y)$ is the cost of moving sand from point x to y (a natural choice is $c(x, y) = |x - y|$); f is the density of the original pile of sand, i.e., $\int_A f(x) \, \mathrm{d}x$ is the mass of sand occupying the set A in the original pile; g is the density of the target distribution (the embankment), i.e., $\int_B g(y) \, \mathrm{d}y$ is the mass of sand occupying the set B in the embankment; $\int_X f(x) \, \mathrm{d}x = \int_Y g(y) \, \mathrm{d}y = 1$ is the total mass of sand (normalised without loss of generality to be 1); T is the transport map - sand at point x in the original pile is transported to point $T(x)$ in the embankment; and the total cost of moving the sand is $M(T)$. The constraint $T\#f = g$ represents conservation of mass - no sand is created or lost in the transportation process:

$$\int_{T^{-1}(B)} f(x) \, \mathrm{d}x = \int_B g(y) \, \mathrm{d}y \quad \forall \, B \subseteq Y$$

which means that the mass of sand transported from the original pile to B equals the mass of the sand in the embankment at B.

Remark 3.3 (The Monge problem for more general densities)**.** The assumptions $\int_X f(x) \, \mathrm{d}x = 1$, $\int_Y g(y) \, \mathrm{d}y = 1$ are not strictly necessary. The Monge problem can also be defined if f and g simply have the same total mass, not necessarily equal to 1: $\int_X f(x) \, \mathrm{d}x = \int_Y g(y) \, \mathrm{d}y$. If $\int_X f(x) \, \mathrm{d}x \neq \int_Y g(y) \, \mathrm{d}y$, then there does not exist any admissible map T satisfying $T\#f = g$ and so $\mathcal{T}_c(f, g) = +\infty$.

The following form of the push-forward constraint is often more useful for calculations:

Lemma 3.4 (Equivalent formulation of the push-forward constraint)**.** *Let $X, Y \subseteq \mathbb{R}^d$ and $T : X \to Y$. Let f be a probability density on X and g be a probability density on Y. Then $T\#f = g$ if and only if*

$$\int_Y \varphi(y) g(y) \, \mathrm{d}y = \int_X \varphi(T(x)) f(x) \, \mathrm{d}x \tag{3.1}$$

for all bounded functions $\varphi : Y \to \mathbb{R}$.

Proof. Suppose that equation (3.1) holds. Let $B \subseteq \mathbb{R}^d$ and choose $\varphi = \chi_B$. Then (3.1) reduces to (2.1) and so $T\#f = g$.

We just sketch the proof of the other direction. Suppose that $T\#f = g$. Then equation (3.1) holds for all characteristic functions $\varphi = \chi_B$, $B \subseteq Y$. By linearity of the integral, it also holds for all simple functions of the form $\varphi = \sum_{i=1}^N a_i \chi_{B_i}$, $a_i \in \mathbb{R}$, $B_i \subseteq Y$. For general bounded functions $\varphi : Y \to \mathbb{R}$, equation (3.1) can be proved by approximating φ by a sequence of simple functions. □

Exercise 3.5. Define $T : \mathbb{R} \to \mathbb{R}$ by $T(x) = 2 - x$. Let $f = \chi_{[0,1]}$ and $g = \chi_{[1,2]}$. Use Lemma 3.4 to show that $T\#f = g$.

The following example is very simple, but it illustrates the complexity of the Monge problem.

Example 3.6. Let $X = [0, 1]$, $Y = [1, 2]$. Let $f(x) = 1$, $x \in X$, and $g(y) = 1$, $y \in Y$. We compare the transport cost of three different transport maps. Let T_1 be the translation $T_1(x) = x + 1$, which transports all the mass the same distance, 1. Let $T_2(x) = 2 - x$, which flips or reflects the mass about the point $x = 1$. The point $x = 1$ is transported distance 0 while the point $x = 0$ is transported distance 2. We could also combine translation and flipping, e.g.,

$$T_3(x) = \begin{cases} x + \frac{3}{2} & \text{if } x \in [0, \frac{1}{2}], \\ 2 - x & \text{if } x \in [\frac{1}{2}, 1]. \end{cases}$$

Which of these maps, if any, are optimal? The answer depends of course on the cost c. Let $c(x, y) = h(y - x)$. We will compare the costs $h(s) = s^2$ (which is strictly convex), $h(s) = |s|$ (which is convex but not strictly convex), and $h(s) = |s|^{1/2}$ (which is concave for $s \geq 0$). If $h(s) = s^2$, then

$$\begin{aligned} M(T_1) &= \int_0^1 (T_1(x) - x)^2 f(x)\,\mathrm{d}x = \int_0^1 1^2\,\mathrm{d}x = 1, \\ M(T_2) &= \int_0^1 (T_2(x) - x)^2 f(x)\,\mathrm{d}x = \int_0^1 (2 - 2x)^2\,\mathrm{d}x = \frac{4}{3}, \\ M(T_3) &= \int_0^1 (T_3(x) - x)^2 f(x)\,\mathrm{d}x = \int_0^{1/2} (3/2)^2\,\mathrm{d}x + \int_{1/2}^1 (2 - 2x)^2\,\mathrm{d}x = \frac{31}{24}. \end{aligned}$$

We leave it as an exercise (Exercise 3.7) to check the values in the following table:

$h(s)$	$M(T_1)$	$M(T_2)$	$M(T_3)$
s^2	1	$\frac{4}{3} \approx 1.33$	$\frac{31}{24} \approx 1.29$
$\lvert s\rvert$	1	1	1
$\lvert s\rvert^{1/2}$	1	$\frac{2\sqrt{2}}{3} \approx 0.94$	$\frac{\sqrt{3}}{2\sqrt{2}} + \frac{1}{3} \approx 0.95$

For the cost $h(s) = s^2$ we have $M(T_1) < M(T_3) < M(T_2)$. Therefore the translation T_1 is the best map amongst these three maps. In fact we can prove it is the best map amongst all

admissible maps as follows: Let T be any admissible map, $T\#f = g$. Since h is convex, then by Jensen's inequality, Theorem 2.7,

$$\begin{aligned}
M(T) &= \int_0^1 h(T(x) - x) f(x)\,\mathrm{d}x \\
&\geq h\left(\int_0^1 (T(x) - x) f(x)\,\mathrm{d}x\right) \\
&= h\left(\int_0^1 T(x) f(x)\,\mathrm{d}x - \int_0^1 x f(x)\,\mathrm{d}x\right) \\
&= h\left(\int_1^2 y g(y)\,\mathrm{d}y - \int_0^1 x f(x)\,\mathrm{d}x\right) && \text{(by (3.1) with } \varphi(y) = y\text{)} \\
&= h\left(\frac{3}{2} - \frac{1}{2}\right) \\
&= h(1) \\
&= \int_0^1 h(1) f(x)\,\mathrm{d}x \\
&= \int_0^1 h(T_1(x) - x) f(x)\,\mathrm{d}x && \text{(since } T_1(x) - x = 1\text{)} \\
&= M(T_1).
\end{aligned}$$

Therefore $M(T) \geq M(T_1)$ for all admissible transport maps T and so T_1 is an optimal transport map and the optimal transport cost is $\mathcal{T}_c(f, g) = M(T_1) = 1$. In fact it can be shown that T_1 is the *unique* optimal transport map (see Section 5). In the argument above we only used the convexity of h, but not the explicit form of h. Therefore the translation T_1 is an optimal transport map for any convex cost.

For the cost $h(s) = |s|$ we have $M(T_1) = M(T_2) = M(T_3) = 1$. Can we do better than this? The answer is no since the map $h(s) = |s|$ is convex and so the argument above shows that $\mathcal{T}_c(f, g) = M(T_1) = 1$. Therefore all three transport maps T_1, T_2, T_3 are optimal. Surprisingly, it turns out that *any* admissible transport map T is optimal: $M(T) = 1$ for all T; see Exercise 3.9. The lack of uniqueness is due to the lack of strict convexity of h.

Finally, consider the cost $h(s) = |s|^{1/2}$, which is concave for $s \geq 0$. In this case $M(T_2) < M(T_3) < M(T_1)$ and flipping mass is better than translating it. We will see below (Example 4.6) that T_2 is an optimal transport map (in fact it is the unique optimal transport map), whereas T_1 is the *worst* possible transport map (Exercise 3.7).

Exercise 3.7. Check the values in the table in Example 3.6. Use Jensen's inequality to prove that T_1 is the *worst* transport map for the concave cost $h(s) = |s|^{1/2}$.

Exercise 3.8. Let $X = [0, 1]$, $Y = [1, 2]$, $f = \chi_{[0,1]}$, $g = \chi_{[1,2]}$, $c(x, y) = h(|y - x|)$ with $h(s) = (s + 1)\log(s + 1)$, $s \geq 0$. Find an optimal transport map.

Exercise 3.9 (Non-uniqueness for linear costs)**.** This example is taken from [11, Examples 2.14, 2.15]. Let $X, Y \subset \mathbb{R}$ be bounded and $c(x, y) = h(y - x)$ where $h : X \to Y$ is a linear function. Show that *every* admissible transport map is optimal, i.e., show that if $T : X \to Y$, $T\#f = g$, then

$$M(T) = \mathcal{T}_c(f, g).$$

Hint: Compute $M(T)$ and show that it is independent of T.

A similar computation shows the following: If all the mass of f lies to the left of the mass of g, i.e., $\sup\{x : f(x) > 0\} \leq \inf\{y : g(y) > 0\}$, then any admissible transport map is optimal for the cost $c(x, y) = |x - y|$ since $T(x) \geq x$ for all x and so $c(x, T(x)) = |T(x) - x| = T(x) - x = h(T(x) - x)$ for the linear map $h(s) = s$.

Exercise 3.10 (Non-uniqueness for non-strictly convex costs: Book shifting)**.** This example is taken from [11, Example 2.16]. Let $X = [0, 2]$, $Y = [1, 3]$, $f = \frac{1}{2}\chi_{[0,2]}$, $g = \frac{1}{2}\chi_{[1,3]}$, $c(x, y) = h(y - x)$ with $h(s) = |s|$. Let $T_1(x) = x + 1$ and

$$T_2(x) = \begin{cases} x + 2 & \text{if } x \in [0, 1], \\ x & \text{if } x \in (1, 2]. \end{cases}$$

Observe that f and g have mass in common in the interval $[1, 2]$. The map T_2 leaves the common mass fixed and only transports mass from $[0, 1]$ to $[2, 3]$. Show that T_1 and T_2 are both optimal transport maps:

$$M(T_1) = M(T_2) = \mathcal{T}_c(f, g).$$

In fact it can be shown using Exercise 3.9 that

$$T(x) = \begin{cases} S(x) & \text{if } x \in [0, 1], \\ x & \text{if } x \in (1, 2], \end{cases}$$

is an optimal transport map for any function $S : [0, 1] \to [2, 3]$ such that $S\#\chi_{[0,1]} = \chi_{[2,3]}$.

Example 3.11 (Non-existence for a strictly concave cost with overlapping masses)**.** Let $X = [0, 1]$, $Y = [0, 2]$, $f = \chi_{[0,1]}$, $g = \frac{1}{2}\chi_{[0,2]}$, $c(x, y) = |x - y|^{1/2}$. In this case it can be shown that there does not exist any optimal transport map; the infimum in the definition of $\mathcal{T}_c(f, g)$ is not attained. It turns out that

$$\mathcal{T}_c(f, g) = \mathcal{T}_c\left(\tfrac{1}{2}\chi_{[0,1]}, \tfrac{1}{2}\chi_{[1,2]}\right).$$

In other words, the cost of transporting f to g is the same as the cost of transporting $\frac{1}{2}\chi_{[0,1]}$ to $\frac{1}{2}\chi_{[1,2]}$. The problem here is that c is strictly concave and the masses f and g 'overlap'; f and g are both positive on the interval $[0, 1]$. Whenever c is a strictly concave metric, as it is here (any function $c(x, y) = h(|y - x|)$ with $h : [0, \infty) \to \mathbb{R}$ concave and $h(0) = 0$ is a metric), then it turns out that it is best to leave 'common mass' where it is. But since any function must take a single value at every point, it is not possible to find a function T that both leaves the common mass $\frac{1}{2}\chi_{[0,1]}$ fixed ($T(x) = x$ on $[0, 1]$) and transports the mass $\frac{1}{2}\chi_{[0,1]}$ to $\frac{1}{2}\chi_{[1,2]}$ ($T([0, 1]) = [1, 2]$).

Exercise 3.12 (A challenging exercise: Behaviour of quadratic transport under translations)**.** This example is taken from [8, Remark 2.19]. Let $X = Y = \mathbb{R}$ and c be the quadratic cost $c(x, y) = (x - y)^2$. For $a \in \mathbb{R}$, define the translation $\tau_a : \mathbb{R} \to \mathbb{R}$ by $\tau_a(x) = x - a$. Let $f \circ \tau_a$ denote the composition $(f \circ \tau_a)(x) = f(\tau_a(x)) = f(x - a)$. In this exercise we show that

$$\mathcal{T}_c(f \circ \tau_a, g \circ \tau_b) = \mathcal{T}_c(f, g) + (b - a)^2 + 2(b - a)(m_g - m_f) \tag{3.2}$$

where $a, b \in \mathbb{R}$ and

$$m_f = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x, \quad m_g = \int_{-\infty}^{\infty} y g(y)\,\mathrm{d}y$$

are the centres of mass of f and g.

(i) Let $T\#f = g$. Define $S : \mathbb{R} \to \mathbb{R}$ by $S(x) = T(x-a)+b$. Show that $S\#(f \circ \tau_a) = g \circ \tau_b$.

(ii) Show that

$$\mathcal{T}_c(f \circ \tau_a, g \circ \tau_b) \leq \mathcal{T}_c(f,g) + (b-a)^2 + 2(b-a)(m_g - m_f).$$

Hint: Let T be an optimal transport map transporting f to g, which means that $\mathcal{T}_c(f,g) = \int_{-\infty}^{\infty} |T(x)-x|^2 f(x)\,\mathrm{d}x$. By part (i),

$$\mathcal{T}_c(f \circ \tau_a, g \circ \tau_b) \leq \int_{-\infty}^{\infty} |S(x)-x|^2 f(\tau_a(x))\,\mathrm{d}x.$$

(iii) Use a similar argument to part (ii) to show that

$$\mathcal{T}_c(f \circ \tau_a, g \circ \tau_b) \geq \mathcal{T}_c(f,g) + (b-a)^2 + 2(b-a)(m_g - m_f).$$

Combining (ii) and (iii) proves (3.2). Hint: Start with an optimal map T transporting $f \circ \tau_a$ to $g \circ \tau_b$. Use it to construct an admissible map S transporting f to g.

(iv) Use (3.2) to give an alternative proof that $\mathcal{T}_c(\chi_{[0,1]}, \chi_{[1,2]}) = 1$.

4 The Dual Problem

In this section we will see that the Monge minimisation problem can be reformulated as a maximisation problem. This is not just a mathematical curiosity, it was an important step along the road to solving Monge's problem.

Throughout this section we assume that $X, Y \subset \mathbb{R}^d$ are compact (closed and bounded). Let $C(X)$ denote the set of continuous, real-valued functions on X.

Theorem 4.1 (Kantorovich Duality Theorem)**.** *Let f be a probability density on X, g be a probability density on Y, and $c : X \times Y \to \mathbb{R}$ be continuous. Define $D : C(X) \times C(Y) \to \mathbb{R}$ by*

$$D(\phi,\psi) = \int_X \phi(x) f(x)\,\mathrm{d}x + \int_Y \psi(y) g(y)\,\mathrm{d}y.$$

If $\phi \in C(X)$, $\psi \in C(Y)$, we say that $\phi \oplus \psi \leq c$ if and only if $\phi(x) + \psi(y) \leq c(x,y)$ for all $x \in X, y \in Y$. Then

$$\mathcal{T}_c(f,g) = \inf_{T\#f=g} M(T) = \sup_{\phi\oplus\psi\leq c} D(\phi,\psi).$$

Moreover, the supremum is a maximum, i.e., there exists an admissible pair (ϕ,ψ) such that $\mathcal{T}_c(f,g) = D(\phi,\psi)$, and we say that (ϕ,ψ) is an optimal Kantorovich potential pair.

Proof. The proof of this theorem is beyond the scope of this short course. We limit ourselves to the proof of the 'easy half' of the duality equality:

$$\inf_{T\#f=g} M(T) \geq \sup_{\phi\oplus\psi\leq c} D(\phi,\psi).$$

Let T satisfy $T\#f = g$ and (ϕ, ψ) satisfy $\phi \oplus \psi \leq c$. Then

$$\begin{aligned} D(\phi,\psi) &= \int_X \phi(x) f(x)\,\mathrm{d}x + \int_Y \psi(y) g(y)\,\mathrm{d}y \\ &= \int_X \phi(x) f(x)\,\mathrm{d}x + \int_X \psi(T(x)) f(x)\,\mathrm{d}x \qquad \text{(by (3.1) with } \varphi = \psi) \\ &= \int_X (\phi(x) + \psi(T(x))) f(x)\,\mathrm{d}x \\ &\leq \int_X c(x, T(x)) f(x)\,\mathrm{d}x \qquad \text{(since } \phi \oplus \psi \leq c) \\ &= M(T). \end{aligned}$$

We record for future use that

$$D(\phi,\psi) = \int_X (\phi(x) + \psi(T(x))) f(x)\,\mathrm{d}x \leq M(T). \tag{4.1}$$

In particular

$$D(\phi,\psi) \leq M(T)$$

for all T satisfying $T\#f = g$ and all (ϕ, ψ) satisfying $\phi \oplus \psi \leq c$. Taking the supremum over all admissible (ϕ, ψ) gives

$$\sup_{\phi\oplus\psi\leq c} D(\phi,\psi) \leq M(T).$$

Then taking the infimum over all admissible T gives

$$\sup_{\phi\oplus\psi\leq c} D(\phi,\psi) \leq \inf_{T\#f=g} M(T)$$

as required. □

Remark 4.2 (The constraint $\phi\oplus\psi \leq c$)**.** By examining the proof of Theorem 4.1 more closely we see that the constraint $\phi(x) + \psi(y) \leq c(x, y)$ does not need to hold for all $x \in X, y \in Y$, but only for x, y such that $f > 0$ in a neighbourhood of x and $g > 0$ in a neighbourhood of y.

Exercise 4.3 (Non-uniqueness of optimal Kantorovich potential pairs)**.** Show that if $\phi\oplus\psi \leq c$ and $a \in \mathbb{R}$, then $(\phi + a) \oplus (\psi - a) \leq c$ and $D(\phi + a, \psi - a) = D(\phi, \psi)$. Therefore if (ϕ, ψ) is an optimal Kantorovich potential pair, then so is $(\phi + a, \psi - a)$ for any $a \in \mathbb{R}$.

Let $\nabla_x c$ denote the gradient of c with respect to its first argument. The following is a useful corollary of the Kantorovich Duality Theorem:

Corollary 4.4. *Let T be a continuous optimal transport map and (ϕ, ψ) be a differentiable optimal Kantorovich potential pair, in particular $M(T) = D(\phi, \psi) = \mathcal{T}_c(f, g)$. Assume that c is differentiable in its first argument. Then*

$$\phi(x_0) + \psi(T(x_0)) = c(x_0, T(x_0)), \qquad \nabla\phi(x_0) = \nabla_x c(x_0, T(x_0))$$

for all $x_0 \in X$ such that $f > 0$ in a neighbourhood of x_0. In particular, we have equality in the inequality constraint $\phi(x) + \psi(y) \leq c(x, y)$ if mass is transported from x to $y = T(x)$.

Proof. Suppose for contradiction that we can find $x_0 \in X$, $r > 0$ such that $f > 0$ on the ball $B_r(x_0)$ and $\phi(x_0) + \psi(T(x_0)) < c(x_0, T(x_0))$. Then by continuity of ϕ, ψ, T, c we have $\phi(x) + \psi(T(x)) < c(x, T(x))$ in some neighbourhood of x_0 where $f > 0$. Combining this with equation (4.1) gives

$$D(\phi, \psi) = \int_X (\phi(x) + \psi(T(x)))f(x)\,\mathrm{d}x < \int_X c(x, T(x))f(x)\,\mathrm{d}x = M(T).$$

But this contradicts the optimality of T and (ϕ, ψ).

Take any $x_0 \in X$ such that $f > 0$ in a neighbourhood of x_0. Consider the map $F(x) = c(x, T(x_0)) - \phi(x)$, $x \in X$. We have shown that

$$F(x) \geq \psi(T(x_0)) \;\forall\, x \in X \quad (\text{since } \phi \oplus \psi \leq c), \qquad F(x_0) = \psi(T(x_0)).$$

Therefore x_0 is a minimum point of F and so

$$\nabla F(x_0) = 0 \quad \iff \quad \nabla_x c(x_0, T(x_0)) - \nabla\phi(x_0) = 0$$

as required. □

We have seen that one optimisation problem, $\inf M(T)$, can be replaced by another, $\sup D(\phi, \psi)$. Why is this useful?

Remark 4.5 (Why the dual problem is useful.)**.** Given a transport map T, how do we know if it is optimal? The Kantorovich Duality Theorem gives us a way of checking. Given any admissible transport map T and admissible Kantorovich potential pair (ϕ, ψ) we have

$$D(\phi, \psi) \leq \mathcal{T}_c(f, g) \leq M(T)$$

by Theorem 4.1. Therefore if we can construct T and (ϕ, ψ) such that $M(T) = D(\phi, \psi)$, then

$$D(\phi, \psi) \leq \mathcal{T}_c(f, g) \leq M(T) = D(\phi, \psi) \quad \implies \quad D(\phi, \psi) = \mathcal{T}_c(f, g) = M(T)$$

and so T and (ϕ, ψ) must be optimal. Even if we can't construct such a T and (ϕ, ψ), then the *duality gap* $M(T) - D(\phi, \psi) \geq 0$ gives us an idea of how far T and (ϕ, ψ) are from being optimal.

Example 4.6. Let $X = [0, 1]$, $Y = [1, 2]$, $f = \chi_{[0,1]}$, $g = \chi_{[1,2]}$. The following table gives optimal Kantorovich potentials and transport maps for the costs from Example 3.6:

$h(s)$	$T(x)$	$\phi(x)$	$\psi(y)$	$D(\phi, \psi) = M(T)$
s^2	$T_1(x) = x + 1$	$-2x$	$2y - 1$	1
$\lvert s\rvert$	$T_1(x) = x + 1$	$-x$	y	1
$\lvert s\rvert^{1/2}$	$T_2(x) = 2 - x$	$\frac{1}{2}(2 - 2x)^{1/2}$	$\frac{1}{2}(2y - 2)^{1/2}$	$\frac{2\sqrt{2}}{3}$

For example, for the cost $h(s) = s^2$, then

$$D(-2x, 2y - 1) = \int_0^1 (-2x)\,\mathrm{d}x + \int_1^2 (2y - 1)\,\mathrm{d}y = -1 + \frac{1}{4}(9 - 1) = 1.$$

Therefore $D(-2x, 2y-1) = M(x+1) = \mathcal{T}_c(f,g) = 1$, which again verifies that the translation $T_1(x) = x+1$ is an optimal transport map for the quadratic cost. Where did the potentials $(\phi(x), \psi(y)) = (-2x, 2y-1)$ come from? We can derive them as follows: By Corollary 4.4, if T_1 and (ϕ, ψ) are optimal, then

$$\phi'(x) = c_x(x, T_1(x)) = 2(x - T_1(x)) = -2 \quad \text{for } x \in [0,1].$$

Integrating gives $\phi(x) = -2x + a$, $x \in [0,1]$. We can choose $a = 0$ by Exercise 4.3. Using Corollary 4.4 again (and again assuming that T_1 is optimal) gives

$$\psi(T_1(x)) = c(x, T_1(x)) - \phi(x) = (T_1(x) - x)^2 + 2x = 1 + 2x \quad \text{for } x \in [0,1].$$

By setting $y = T_1(x) = x + 1$ we find that

$$\psi(y) = 1 + 2(y-1) = 2y - 1 \quad \text{for } y \in [1,2].$$

It is an exercise in multivariable calculus to verify that

$$\phi \oplus \psi \le c \quad \Longleftrightarrow \quad -2x + 2y - 1 \le (y-x)^2 \quad \forall\, (x,y) \in X \times Y.$$

Hint: Find the maximum value of $F(x,y) = -2x + 2y - 1 - (y-x)^2$ in $\mathbb{R}^2$.

We leave it as an exercise (Exercise 4.8) to check the other values in the table above or, even better, to derive the optimal Kantorovich potential pairs for yourself.

Remark 4.7 (Kantorovich potentials for metric costs)**.** Consider the special case where $X = Y$ and c is a metric on X, e.g., $X = \mathbb{R}^d$ and $c(x,y) = |y-x|$ or $c(x,y) = h(|y-x|)$ for a concave function $h : [0,\infty) \to [0,\infty)$ with $h(0) = 0$. In this case it can be shown (with a lot of work) that there exists an optimal Kantorovich potential pair (ϕ, ψ) with $\psi = -\phi$ and where ϕ is 1-Lipschitz with respect to c, which means that $|\phi(x) - \phi(y)| \le c(x,y)$ for all $x, y \in X$. For instance, if $c(x,y) = |y-x|$, then we can choose optimal Kantorovich potentials (ϕ, ψ) such that $\psi(y) = -\phi(y)$ and $|\phi(x) - \phi(y)| \le |x-y|$.

Exercise 4.8. Fill in the missing details for Example 4.6.

Exercise 4.9. Derive an optimal Kantorovich potential pair for the book shifting problem from Exercise 3.10.

Exercise 4.10. Prove that T_2 is the *worst* transport map for the convex cost $h(s) = s^2$ from Example 3.6. Hint: This is equivalent to proving that T_2 is the best transport map for the concave cost $\tilde{h}(s) = -s^2$. Verify this by constructing an optimal Kantorovich potential pair (ϕ, ψ) such that $D(\phi, \psi) = M(T_2)$ for the cost $\tilde{h}(s) = -s^2$.

5 Fundamental Theorems of Optimal Transport Theory

We finish the course by stating some of the big theorems of optimal transport theory, which came over 200 years after Gaspard Monge posed the optimal transport problem.

First note that the definition of convex functions, Definition 2.2, can be extended to multivariable functions: We say $h : \mathbb{R}^d \to \mathbb{R}$ is convex if for all $\lambda \in (0,1)$, $x, y \in \mathbb{R}^d$

$$h((1-\lambda)x + \lambda y) \le (1-\lambda)h(x) + \lambda h(y).$$

The first breakthrough came for the quadratic cost:

Theorem 5.1 (Brenier's Theorem [3])**.** *Let $X = Y = \mathbb{R}^d$ and f, g be probability densities on $\mathbb{R}^d$ that vanish outside a compact set. Let c be the quadratic cost $c(x, y) = (x-y)^2$. Then there exists a unique optimal transport map T. Moreover, T is the gradient of a convex function, i.e., $T = \nabla\Phi$ for some convex function Φ. If $d = 1$ and T is differentiable, then $T = \Phi'$ and so $T' = \Phi'' \geq 0$ (since Φ is convex). Therefore T is a non-decreasing function.*

Note that *unique* in Brenier's Theorem means that T is unique on the set where f has positive mass.

Example 5.2. For the quadratic cost $h(s) = s^2$ in Example 3.6, we found the optimal transport map $T_1(x) = x + 1 = \Phi'(x)$ for the convex function $\Phi(x) = \frac{1}{2}(x+1)^2$.

Brenier's Theorem is actually a (nontrivial) consequence of the following theorem, which came a few years later:

Theorem 5.3 (Strictly convex costs [6])**.** *Let $X = Y = \mathbb{R}^d$ and f, g be probability densities on $\mathbb{R}^d$ that vanish outside a compact set. Let c be the cost $c(x, y) = h(x - y)$ where h is strictly convex. Then there exists a unique optimal transport map T. If (ϕ, ψ) is an optimal Kantorovich potential pair, then T and ϕ are related by*

$$T(x) = x - (\nabla h)^{-1}(\nabla\phi(x))$$

for all x (except for a set of mass 0 with respect to f). The potential ϕ is not unique, but its gradient $\nabla\phi$ is.

Proof. The proof of this theorem is well beyond the scope of this course. The form of T, however, can be inferred from Corollary 4.4:

$$\nabla\phi(x) = \nabla_x c(x, T(x)) = \nabla h(x - T(x)) \quad \Longrightarrow \quad T(x) = x - (\nabla h)^{-1}(\nabla\phi(x)).$$

□

Example 5.4. Let's verify the formula $T(x) = x - (\nabla h)^{-1}(\nabla\phi(x))$ for Example 4.6. In one dimension this reduces to $T(x) = x - (h')^{-1}(\phi'(x))$. For the quadratic cost $h(s) = s^2$ we found $T(x) = x + 1$ and $\phi(x) = -2x$. We have $h'(s) = 2s$, $(h')^{-1} = s/2$, and

$$x - (h')^{-1}(\phi'(x)) = x - \frac{\phi'(x)}{2} = x + 1 = T(x)$$

as required.

Remark 5.5 (Strictly convex costs in one dimension)**.** One dimension is very special; for strictly convex costs there is an explicit formula for the optimal transport map T, T is non-decreasing and, surprisingly, it is independent of the cost c [11, Theorem 2.9]:

$$T(x) = F_g^{[-1]}(F_f(x))$$

where F_f and F_g are the cumulative distribution functions

$$F_f(t) = \int_{-\infty}^{t} f(x)\,\mathrm{d}x, \qquad F_g(t) = \int_{-\infty}^{t} g(y)\,\mathrm{d}y$$

and $F_g^{[-1]}$ is the pseudo-inverse of F_g:

$$F_g^{[-1]}(x) = \inf\{t \in \mathbb{R} : F_g(t) > x\}.$$

This agrees with Example 3.6, where we showed that T_1 is optimal for any convex cost.

Remark 5.6 (The cost $c(x,y) = |x-y|$)**.** Unfortunately Theorem 5.3 does not include what is possibly the the most physical cost, $c(x-y) = h(x-y)$ with $h(s) = |s|$, which is convex but not strictly convex. In this case there still exists an optimal transport map T (this was established by [1, 4]) but, as we have seen, T is not necessarily unique.

Remark 5.7 (Strictly concave costs)**.** If h is strictly concave then there exists a unique optimal transport map T provided that f and g do not share any mass, i.e., for any $A \subset \mathbb{R}^d$ such that $\int_A f(x)\,dx > 0$, then $\int_A g(y)\,dy = 0$ [12, Theorem 2.45]. If f and g share mass, then there may not exist an optimal transport map; see Example 3.11.

References

[1] L. Ambrosio. Lecture notes on optimal transport problems. In *Mathematical Aspects of Evolving Interfaces*, volume 1812 of *Lecture Notes in Mathematics*, pages 1–52. Springer, 2003.

[2] L. Ambrosio and N. Gigli. A user's guide to optimal transport. In *Modelling and Optimisation of Flows on Networks*, volume 2062 of *Lecture Notes in Mathematics*, pages 1–155. Springer, 2013.

[3] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Commun. Pure Appl. Math.*, 44:375–417, 1991.

[4] L. C. Evans and W. Gangbo. Differential equations methods for the Monge-Kantorovich mass transfer problem. *Mem. Am. Math. Soc.*, 137, 1999.

[5] A. Galichon. *Optimal Transport Methods in Economics.* Princeton University Press, 2016.

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

[7] G. Monge. Mémoire sur la théorie des déblais et des remblais. In *Histoire de l'Académie Royale des Sciences de Paris*, pages 666–704, 1781.

[8] G. Peyré and M. Cuturi. Computational optimal transport. arXiv:1803.00567, 2018.

[9] S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems, Vol. I: Theory.* Sringer, 1998.

[10] S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems, Vol. II: Applications.* Sringer, 1998.

[11] F. Santambrogio. *Optimal Transport for Applied Mathematicians.* Birkhäuser, 2015.

[12] C. Villani. *Topics in Optimal Transportation.* American Mathematical Society, 2003.

[13] C. Villani. *Optimal Transport: Old and New.* Springer, 2009.