TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

D I P L O M A R B E I T

Optimal Transport on the n-Sphere

Ausgeführt am Institut für

Diskrete Mathematik und Geometrie
der Technischen Universität Wien

unter Anleitung von

Assoc. Prof. Dr. Franz Schuster

und Mitwirkung von

Univ.Ass. Dipl.-Ing. Dr.techn. Gabriel Maresch

durch

Florian Besau, BSc
Pfadenhauergasse 2/10
1140 Wien

Datum

Unterschrift

Preface

Optimal Transport was first introduced by Gaspard Monge in 1781. In the 1940s Leonid Kantorovich relaxed the original problem and introduced a dual representation [Kan42, Kan48]. Since then many great mathematicians have contributed to this field and made it a classic topic in probability and optimization theory.

In this thesis I aim to "explicitly" solve the Optimal Transportation problem on the n-dimensional sphere.

In Chapter 1, I will give a quick survey of convex functions. The theory of convex functions, and in particular the dual correspondence via conjugate functions, will prove useful when we establish the Kantorovich duality in Chapter 2. Besides basic knowledge in analysis and linear algebra no further knowledge is presupposed. Most of the results and proofs in this chapter are taken from [Roc70].

In Chapter 2, I will motivate the definition of the original Monge problem (MP) and its relaxation, the Monge-Kantorovich problem (MKP). I will prove the Kantorovich duality and introduce c-concave functions and the c-transform of such functions. In the final part of Chapter 2, I will solve the Optimal Transport problem on the real line for convex cost functions. Basic knowledge in measure theory is recommended. The content of this chapter is mostly along the lines of Villani's work in [Vil03] and [Vil09].

In Chapter 3, I will study Optimal Transport on the n-dimensional sphere. I will first use our results from Chapter 2 on the real line to solve the MKP on the circle for convex cost functions. Then, I will prove the main theorem of Optimal Transport on the n-sphere. I will introduce differential calculus and a version of Rademacher's Theorem on the n-sphere. In the final part of this chapter I will use these tools together with the Kantorovich duality to solve the Optimal Transport problem "explicitly" on the n-sphere. My strategy will be similar to McCann's strategy in [McC01], where he treats the Optimal Transport problem on Riemannian manifolds. In this chapter the reader will require basic knowledge in geometry, but no knowledge about Riemannian manifolds.

Acknowledgments

I would like to thank Professor Dr. Monika Ludwig and Dr. Gabriel Maresch who first sparked my interest in the topic of optimal transportation in a seminar back in the summer of 2011. Gabriel initiated this thesis afterwards in the fall of 2011 and has been providing me with valuable advice and comments ever since, for which I am very grateful. My gratitude also belongs to Associate Professor Dr. Franz Schuster, who joined later on and has been an extremely welcome support – I am very excited to work with him in the future!

I also wish to thank a number of professors for their lectures at the university and my fellow students for various mathematical and non-mathematical discussions as well as their much appreciated companionship.

My gratitude also goes to my family, close relatives and friends who supported me wholeheartedly and were quick to help when my mind needed distraction.

Finally, I would like to thank the proofreaders of this thesis: Astrid Berg, Patrick Heimel, Patrick Marschik, Michael Vögler and Sebastian Zivota. I am very grateful for their effort, which most certainly improved this thesis.

Florian Besau

Contents

Chapter 1

Convex Geometry

In this chapter we recall some results from convex geometry on convex sets and functions. Most of the following results are taken from [Roc70] and [Sch93], which we recommend for a more thorough study of this fascinating topic.

1.1 Convex Sets and Functions

Definition 1.1.1 (Convex set)**.** A non-empty subset $C \subseteq \mathbb{R}^n$ is called **convex** if for all $x, y \in C$

$$\lambda x + (1-\lambda)y \in C, \qquad \forall \lambda \in (0,1). \tag{1.1}$$

Example 1.1.2. Some basic examples for convex sets are

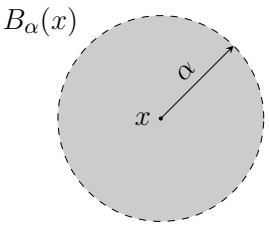
(a) **Affine sets**: A set $E \subseteq \mathbb{R}^n$ is called **affine** if for all $x, y \in E$ and $\lambda \in \mathbb{R}$ we have

$$(1-\lambda)x + \lambda y \in E.$$

(b) **Balls**: Let $\alpha > 0$ and $x \in \mathbb{R}^n$. We define the **open ball** centered in x of radius α by

$$B_\alpha(x) := \{y \in \mathbb{R}^n \,|\, \|x-y\| < \alpha\},$$

where $\|.\| : \mathbb{R}^n \to [0, +\infty)$ denotes the **Euclidean norm**. Also, we denote the closure of $B_\alpha(x)$ by $\overline{B_\alpha}(x)$.



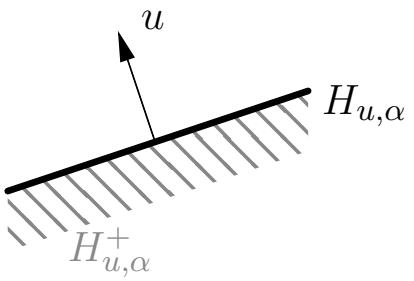
(c) **Hyperplanes and half-spaces**: Let $u \in \mathbb{R}^n \backslash \{0\}$ and $\alpha \in \mathbb{R}$. The set $H_{u,\alpha}$ defined by

$$H_{u,\alpha} := \{y \in \mathbb{R}^n | \langle y, u \rangle = \alpha\},$$

is affine and is called a **hyperplane**, where

$$\langle ., . \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$$

denotes the **Euclidean scalar product**.



The set $H^+_{u,\alpha}$ defined by

$$H^+_{u,\alpha} := \{y \in \mathbb{R}^n | \langle y, u \rangle < \alpha\},$$

is called an open **half-space**. The closure of $H^+_{u,\alpha}$ is denoted by $\overline{H}^+_{u,\alpha}$. We have

$$\overline{H}^+_{u,\alpha} = H_{u,\alpha} \cup H^+_{u,\alpha}.$$

Also, we define $H^-_{u,\alpha}$ by

$$H^-_{u,\alpha} := \{y \in \mathbb{R}^n | \langle y, u \rangle > \alpha\},$$

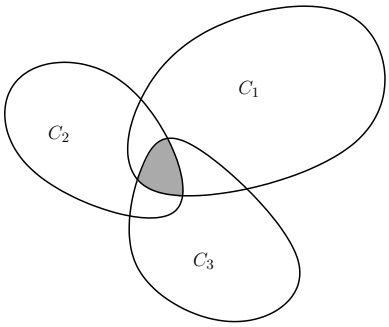
and denote its closure by $\overline{H}^-_{u,\alpha}$.

Convexity of a set is preserved under various operations.

Proposition 1.1.3.

(i) Let $(C_i)_{i \in I}$ be a family of convex sets. Then

$$\bigcap_{i \in I} C_i,$$



is convex if it is non-empty.

(ii) Let C_1, C_2 be convex sets, then

$$C_1 + C_2 := \{x + y | x \in C_1, y \in C_2\}$$

is convex.

(iii) Let $A : \mathbb{R}^m \to \mathbb{R}^n$ be a linear transformation and $C \subseteq \mathbb{R}^m$, $D \subseteq \mathbb{R}^n$ be convex sets, then $AC \subseteq \mathbb{R}^n$ and $A^{-1}D \subseteq \mathbb{R}^m$ are convex sets.

Proof.

ad (i): Let x, $y \in C$, then necessarily $x \in C_i$ and $y \in C_i$ for all $i \in I$. Thus, for $\lambda \in (0,1)$, we have

$$\lambda x + (1-\lambda)y \in C_i, \qquad \forall i \in I.$$

Hence $\lambda x + (1-\lambda)y \in C$ and therefore C is convex.

ad (ii): Let x, $y \in C_1 + C_2$, then there are x_1, $y_1 \in C_1$ and x_2, $y_2 \in C_2$, such that

$$x = x_1 + x_2 \text{ and } y = y_1 + y_2.$$

For $\lambda \in (0,1)$, we get

$$\lambda x_1 + (1-\lambda)x_2 \in C_1, \qquad \lambda y_1 + (1-\lambda)y_2 \in C_2$$

and therefore

$$\lambda x + (1-\lambda)y \in C_1 + C_2,$$

thus $C_1 + C_2$ is convex.

ad (iii): Let x, $y \in AC$. There are u, $v \in C$, such that $x = Au$ and $y = Av$. For $\lambda \in (0,1)$, we have

$$\lambda u + (1-\lambda)v \in C.$$

Thus, since A is linear, we get

$$\lambda x + (1-\lambda)y = \lambda Au + (1-\lambda)Av = A(\lambda u + (1-\lambda)v) \in AC.$$

Therefore AC is convex.

Now, let x, $y \in A^{-1}D$, then Ax, $Ay \in D$. For $\lambda \in (0,1)$, we have

$$A(\lambda x + (1-\lambda)y) = \lambda Ax + (1-\lambda)Ay \in D.$$

Thus

$$\lambda x + (1-\lambda)y \in A^{-1}D$$

and therefore $A^{-1}D$ is convex. □

Remark 1.1.4. Let $\lambda > 0$ and $\mu > 0$. For $B \subseteq \mathbb{R}^n$, we have

$$(\lambda + \mu)B \subseteq \lambda B + \mu B.$$

If $A \subseteq \mathbb{R}^n$ is convex, we have

$$\tfrac{\lambda}{\lambda+\mu}A + \tfrac{\mu}{\lambda+\mu}A \subseteq A,$$

thus we have

$$\lambda A + \mu A = (\lambda + \mu)A. \tag{1.2}$$

Equation (1.2) characterizes convex sets: A is convex if and only if $A \neq \emptyset$ and (1.2) holds for all $\lambda > 0$ and $\mu > 0$.

Theorem 1.1.5. *Let $C \subseteq \mathbb{R}^n$ be convex and closed. Then the intersection of all closed half-spaces containing C is equal to C, i.e.*

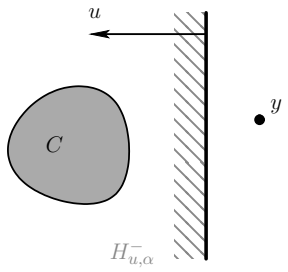
$$C = \bigcap \{\overline{H} |\ C \text{ is contained in the closed half-space } \overline{H}\}.$$

Proof. Obviously we have

$$C \subseteq \bigcap \{\overline{H} |\ C \text{ is contained in the closed half-space } \overline{H}\}.$$

If $C = \mathbb{R}^n$ there is nothing to show. Let $y \notin C$, then $\{y\}$ is compact and C is closed. Thus, by the **Hahn-Banach separation Theorem**, there is $u \in \mathbb{R}^n \backslash \{0\}$ and $\alpha \in \mathbb{R}$ such that

$$\langle y, u\rangle < \alpha \leq \inf_{x\in C} \langle x, u\rangle .$$



This yields

$$y \notin \overline{H}^-_{u,\alpha} \supseteq C.$$

So for all $y \notin C$ we find a closed half-space that separates $\{y\}$ and C, thus

$$C \supseteq \bigcap \{\overline{H} |\ C \text{ is contained in the closed half-space } \overline{H}\}.$$

□

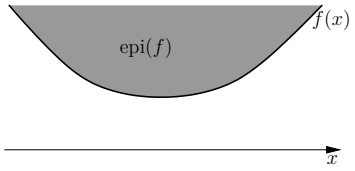
Definition 1.1.6 (Convex function)**.** Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$.

(i) For $\alpha \in \mathbb{R} \cup \{\pm\infty\}$ the set $\{x \in \mathbb{R}^n | f(x) = \alpha\}$ is denoted by $\{f = \alpha\}$. Similar definitions apply for $\{f \leq \alpha\}$, $\{f < \alpha\}$, etc.

(ii) We define the **effective domain** of f by $\mathrm{dom}(f) := \{f < +\infty\}$. f is called **proper** if $\{f = -\infty\} = \emptyset$ and $\mathrm{dom}(f) \neq \emptyset$.

(iii) f is called **convex** if f is proper and the **epigraph** of f, defined by

$$\mathrm{epi}(f) := \{(x, y) \in \mathbb{R}^{n+1} | x \in \mathbb{R}^n, y \in \mathbb{R}, y \geq f(x)\},$$



is a convex set.
f is called **concave** if $-f$ is convex.

(iv) Let $g : A \subseteq \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$. We call g proper if the extension to $\mathbb{R}^n$ by

$$\tilde{g}(x) := \begin{cases} g(x), & \text{if } x \in A \\ +\infty, & \text{otherwise.} \end{cases}$$

is proper.
Also, we call g convex (concave) if $\tilde{g}$ is convex (concave).

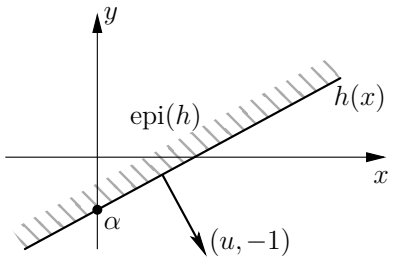
Example 1.1.7 (Affine functions)**.** We call $h : \mathbb{R}^n \to \mathbb{R}$ **affine** if for all x, $y \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$ we have

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

Let $h : \mathbb{R}^n \to \mathbb{R}$ be affine. The function $g : \mathbb{R}^n \to \mathbb{R}$ defined by

$$g(z) := h(z) - h(0)$$

is linear. Thus there is $u \in \mathbb{R}^n$ such that $g(z) = \langle z, u \rangle$. Setting $\alpha = h(0)$ we have



$$h(x) = \langle x, u \rangle - \alpha.$$

This yields

$$\operatorname{epi}(h) = \overline{H}^{+}_{(u,-1),\alpha},$$

hence h is convex.

We note that if f is convex, all sublevel sets $\{f \leq \alpha\}$, $\{f < \alpha\}$ as well as $\operatorname{dom}(f)$ are convex sets. This follows easily from the convexity of $\operatorname{epi}(f)$, because all these sets can be constructed as intersections of $\operatorname{epi}(f)$ with a half-space or projections on $\mathbb{R}^n$ (and by Proposition 1.1.3 these operations preserve convexity).

In the following proposition we will give an equivalent condition for convexity of a function.

Proposition 1.1.8. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper function. f is convex if and only if*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y), \qquad \forall \lambda \in (0,1), \tag{1.3}$$

for every x and y in $\operatorname{dom}(f)$.

Proof. Let f be convex and let x_1, $x_2 \in \operatorname{dom}(f)$. Then $(x_1, f(x_1))$, $(x_2, f(x_2)) \in \operatorname{epi}(f)$ and thus, since $\operatorname{epi}(f)$ is convex, for $\lambda \in (0,1)$ we get

$$\lambda(x_1, f(x_1)) + (1-\lambda)(x_2, f(y_2)) = (\lambda x_1 + (1-\lambda)x_2, \lambda f(x_1) + (1-\lambda)f(x_2)) \in \operatorname{epi}(f).$$

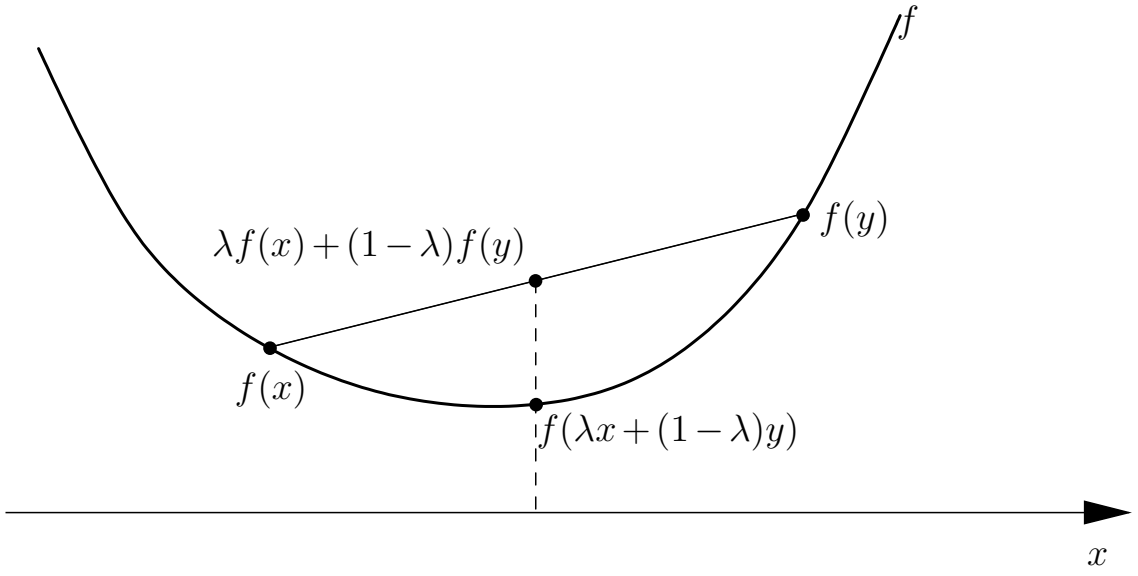


Figure 1.1: Sketch of condition (1.3) in Proposition 1.1.8.

This is equivalent to

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2).$$

Now, assume that f satisfies (1.3) on $\mathrm{dom}(f)$ and let (x_1, y_1), $(x_2, y_2) \in \mathrm{epi}(f)$. Then $y_1 \geq f(x_1)$ and $y_2 \geq f(x_2)$. Therefore

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \leq \lambda y_1 + (1-\lambda) y_2.$$

So $\mathrm{epi}(f)$ is convex and so is f. □

Remark 1.1.9.

1. For a convex function f we can use induction on condition (1.3) to get

$$f\left(\sum_{i=1}^{k} \lambda_i x_i\right) \leq \sum_{i=1}^{k} \lambda_i f(x_i), \tag{1.4}$$

   for all $\lambda_i \geq 0$ $(i = 1, \ldots, n)$ with $\sum_{i=1}^{k} \lambda_i = 1$. This is also known as **Jensen's inequality**.

2. A convex function f is called **strictly convex** if and only if equality in (1.3) implies $x = y$.

As for convex sets, the convexity of a function is preserved under various operations.

Proposition 1.1.10.

(i) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and $g : \mathbb{R} \to \mathbb{R}$ a non-decreasing and convex function. Then $g \circ f$ is convex if $\mathrm{dom}(g \circ f) \neq \emptyset$ (we set $g(+\infty) = +\infty$).

(ii) Let f_1 and f_2 be convex functions on $\mathbb{R}^n$, then $f_1 + f_2$ is convex if $\mathrm{dom}(f_1) \cap \mathrm{dom}(f_2) \neq \emptyset$.

(iii) Let $C \subseteq \mathbb{R}^{n+1}$ be a convex set. Then the function $f : \mathbb{R}^n \to \mathbb{R}$ defined by

$$f(x) := \inf\{\alpha \in \mathbb{R} | (x, \alpha) \in C\},$$

is convex.

(iv) Let f_1, f_2, ..., f_n be convex functions, then

$$(f_1 \square f_2 \square \ldots \square f_n)(x) := \inf\{f_1(x_1) + f_2(x_2) + \ldots + f_n(x_n) | x_1 + x_2 + \ldots + x_n = x\}$$

*is a convex function, called the **infimal convolution** of f_1, f_2, ..., f_n. Also, for $n = 2$, we can write*

$$f_1 \square f_2(x) = \inf_{y \in \mathbb{R}^n} \{f_1(y) + f_2(x - y)\}.$$

(v) Let $(f_i)_{i \in I}$ be a family of convex functions on $\mathbb{R}^n$. Define $f : \mathbb{R}^n \to \mathbb{R}$ by

$$f(x) := \sup_{i \in I} f_i(x).$$

If $\mathrm{dom}(f) \neq \emptyset$, i.e. there exists $x_0 \in \mathbb{R}^n$ such that

$$\sup_{i \in I} f_i(x_0) < +\infty,$$

then f is convex.

(vi) Let $A : \mathbb{R}^m \to \mathbb{R}^n$ and $B : \mathbb{R}^n \to \mathbb{R}^m$ be linear transformations and $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function. Then $fA : \mathbb{R}^m \to \mathbb{R}$ defined by

$$fA(x) := f(Ax)$$

and $Bf : \mathbb{R}^m \to \mathbb{R}$ *defined by*

$$Bf(y) := \inf\{f(x) | Bx = y\}$$

are convex functions.

Proof.

ad (i): First, we notice that $g \circ f$ is proper, since $\mathrm{dom}(g \circ f) \neq \emptyset$.
Let x, $y \in \mathrm{dom}(f)$ and $\lambda \in (0,1)$. We have

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y).$$

Since g is non-decreasing and convex, we get

$$\begin{aligned}(g \circ f)(\lambda x + (1-\lambda)y) &\leq g(\lambda f(x) + (1-\lambda) f(y)) \\ &\leq \lambda (g \circ f)(x) + (1-\lambda)(g \circ f)(y).\end{aligned}$$

Thus $g \circ f$ is convex.

ad (ii): $f_1 + f_2$ is a proper function, since

$$\mathrm{dom}(f_1 + f_2) = \mathrm{dom}(f_1) \cap \mathrm{dom}(f_2) \neq \emptyset.$$

Let x, $y \in \mathrm{dom}(f_1 + f_2)$ and $\lambda \in (0,1)$, then

$$f_i(\lambda x + (1-\lambda)y) \leq \lambda f_i(x) + (1-\lambda) f_i(y)$$

for $i = 1, 2$. Taking the sum of these inequalities we get

$$(f_1 + f_2)(\lambda x + (1-\lambda)y) \leq \lambda (f_1 + f_2)(x) + (1-\lambda)(f_1 + f_2)(y)$$

and therefore $f_1 + f_2$ is convex.

ad (iii): $\mathrm{dom}(f)$ is the projection of C onto $\mathbb{R}^n$. Thus, since C is convex and therefore $C \neq \emptyset$, $\mathrm{dom}(f) \neq \emptyset$ and f is proper.
Let x_1, $x_2 \in \mathrm{dom}(f)$, then, for $\epsilon_1 \geq 0$, $\epsilon_2 \geq 0$ small enough, we have

$$(x_1, f(x_1) + \epsilon_1) \in C \text{ and } (x_2, f(x_2) + \epsilon_2) \in C.$$

Let $\lambda \in (0,1)$. We get

$$\lambda(x_1, f(x_1)+\epsilon_1)+(1-\lambda)(x_2, f(x_2)+\epsilon_2) \in C$$

and therefore

$$f(\lambda x_1+(1-\lambda)x_2) \leq \lambda(f(x_1)+\epsilon_1)+(1-\lambda)(f(x_2)+\epsilon_2).$$

Since ϵ_i can be chosen arbitrarily small, we get

$$f(\lambda x_1+(1-\lambda)x_2) \leq \lambda f(x_1)+(1-\lambda)f(x_2)$$

and therefore f is convex.

ad (iv): Let $F_i := \operatorname{epi}(f_i)$ and $F := F_1 + F_2 + \ldots + F_n$, then F is convex. Further, $(x,y) \in F$ if and only if there exist $(x_i)_{i=1}^n$ in $\mathbb{R}^n$ and $(y_i)_{i=1}^n$ in $\mathbb{R}$, such that $\sum_{i=1}^n x_i = x$, $\sum_{i=1}^n y_i = y$ and $f_i(x_i) \leq y_i$ for all $i = 1, \ldots, n$. Thus

$$f(x) = \inf\{y | (x,y) \in F\}$$

and therefore, by (iii), f is convex.

ad (v): Since $\sup_{i\in I} f_i(x_0) < +\infty$ we have $x_0 \in \operatorname{dom}(f)$ and therefore f is proper. Let x, $y \in \operatorname{dom}(f)$ and $\lambda \in (0,1)$. Since

$$\operatorname{epi}(f) \subseteq \bigcap_{i\in I} \operatorname{epi}(f_i),$$

we have x, $y \in \operatorname{dom}(f_i)$ for all $i \in I$. Thus

$$f_i(\lambda x+(1-\lambda)y) \leq \lambda f_i(x)+(1-\lambda)f_i(y), \qquad \forall i \in I.$$

Therefore

$$\begin{aligned} f(\lambda x+(1-\lambda)y) &= \sup_{i\in I} f_i(\lambda x+(1-\lambda)y) \\ &\leq \sup_{i\in I} \lambda f_i(x)+(1-\lambda)f_i(y) \leq \lambda f(x)+(1-\lambda)f(y), \end{aligned}$$

hence f is convex.

ad (vi): We have $\mathrm{dom}(fA) = A(\mathrm{dom}(f)) \neq \emptyset$, thus fA is proper.
Let x, $y \in \mathrm{dom}(fA)$ and $\lambda \in (0,1)$. Then

$$fA(\lambda x + (1-\lambda)y) = f(\lambda Ax + (1-\lambda)Ay) \leq \lambda fA(x) + (1-\lambda)fA(y)$$

and therefore fA is convex.
Bf is proper, since $\mathrm{dom}(Bf) = B^{-1}\mathrm{dom}(f) \neq \emptyset$.
Let y_1, $y_2 \in \mathrm{dom}(Bf)$. Then, for $\epsilon_1 \geq 0$ and $\epsilon_2 \geq 0$ small enough, there are x_1, $x_2 \in \mathbb{R}^n$ such that $Bx_1 = y_1$, $Bx_2 = y_2$, $Bf(y_1) + \epsilon_1 = f(x_1)$ and $Bf(y_2) + \epsilon_2 = f(x_2)$.
If $\lambda \in (0,1)$, then

$$\begin{aligned} f(\lambda x_1 + (1-\lambda)x_2) &\leq \lambda f(x_1) + (1-\lambda)f(x_2) \\ &= \lambda(Bf(y_1) + \epsilon_1) + (1-\lambda)(Bf(y_2) + \epsilon_2) \end{aligned}$$

and since $B(\lambda x_1 + (1-\lambda)x_2) = \lambda y_1 + (1-\lambda)y_2$ we get

$$\begin{aligned} Bf(\lambda y_1 + (1-\lambda)y_2) &\leq f(\lambda x_1 + (1-\lambda)x_2) \\ &\leq \lambda(Bf(y_1) + \epsilon_1) + (1-\lambda)(Bf(y_2) + \epsilon_2). \end{aligned}$$

Now, since ϵ_i can be chosen arbitrarily small, this implies

$$Bf(\lambda y_1 + (1-\lambda)y_2) \leq \lambda Bf(y_1) + (1-\lambda)Bf(y_2)$$

and therefore Bf is convex. □

1.2 Differential Theory of Convex Functions

In this section we will show that a convex function is locally Lipschitz-continuous. Furthermore, we will show that the directional derivatives exist and are sub-linear. We will then conclude that convex functions are subdifferentiable. We recall definitions when needed.

Definition 1.2.1 (Lipschitz-continuity)**.** Let $f : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$ be a function.

(i) We call f **Lipschitz-continuous** if there exists a constant $B > 0$ such that

$$\|f(x) - f(y)\| \leq B\|x - y\| \ \forall x, y \in A. \tag{1.5}$$

We denote the space of Lipschitz-continuous functions on A by $\mathrm{Lip}(A,\mathbb{R}^m)$. The smallest constant B is given by

$$\mathrm{Lip}(f) := \sup_{x,y\in A: x\neq y} \frac{\|f(x)-f(y)\|}{\|x-y\|}. \tag{1.6}$$

(ii) We call f **locally Lipschitz-continuous** if f is Lipschitz-continuous on all compact $C \subseteq A$.

If f is Lipschitz-continuous on $A \subseteq \mathbb{R}^n$, then it obviously is uniformly continuous as well, because for arbitrary $\epsilon > 0$ we can set $\delta := \frac{\epsilon}{\mathrm{Lip}(f)}$ and get

$$\|f(x)-f(y)\| \leq \mathrm{Lip}(f)\,\|x-y\| < \epsilon,$$

for all x, $y \in A$ such that $\|x-y\| \leq \delta$.

Theorem 1.2.2. *A convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is continuous and locally Lipschitz-continuous on* $\mathrm{int}(\mathrm{dom}(f))$.

Proof. The following proof is taken from Theorem 1.5.1 in [Sch93]. We set $O := \mathrm{int}(\mathrm{dom}(f))$ and assume $O \neq \emptyset$, because otherwise there is nothing to prove. First we will show the continuity of f on O. Let $x_0 \in O$. Since O is non-empty and open, we can choose affinely independent $(x_i)_{i=1}^{n+1}$ in O, such that the simplex S generated by these points satisfies

$$x_0 \in \mathrm{int}(S) \subseteq S \subseteq O.$$

Furthermore, there is $\rho > 0$ such that the open ball $B_\rho(x_0) \subseteq S$. For $x \in S$ there is a representation $x = \sum_{i=1}^{n+1} \lambda_i x_i$ with $\lambda_i \geq 0$ and $\sum_{i=1}^{n+1} \lambda_i = 1$. Using Jensen's inequality (1.4), we deduce that

$$f(x) \leq \sum_{i=1}^{n+1} \lambda_i f(x_i) \leq c := \max_{1\leq i\leq n+1} f(x_i), \qquad \forall x \in S.$$

Now let $y = x_0 + \alpha u$ with $\alpha \in (0,1)$ and $\|u\| = \rho$. Since $y = (1-\alpha)x_0 + \alpha(x_0+u)$, we get

$$f(y) \leq (1-\alpha) f(x_0) + \alpha f(x_0+u),$$

hence

$$f(y)-f(x_0)\leq\alpha(f(x_0+u)-f(x_0))\leq\alpha(c-f(x_0))$$

since $x_0+u\in B_\rho(x_0)\subseteq S$.
On the other hand,

$$x_0=\frac{1}{1+\alpha}y+\frac{\alpha}{1+\alpha}(x_0-u)$$

and hence

$$f(x_0)\leq\frac{1}{1+\alpha}f(y)+\frac{\alpha}{1+\alpha}f(x_0-u),$$

which yields

$$f(x_0)-f(y)\leq\alpha(f(x_0-u)-f(x_0))\leq\alpha(c-f(x_0)).$$

Thus, since $\alpha\rho=\|y-x_0\|$, we get

$$|f(y)-f(x_0)|\leq\alpha(c-f(x_0))=\frac{c-f(x_0)}{\rho}\|y-x_0\|,$$

for $y\in B_\rho(x_0)$. Therefore f is continuous in x_0 and hence, since x_0 was arbitrarily chosen, f is continuous on O.
To prove local Lipschitz-continuity let $C\subseteq O$ be compact. We need to show that f is Lipschitz-continuous on C. By compactness, there exists $\rho>0$ such that

$$C_\rho:=C+B_\rho(0)\subseteq O.$$

On the compact set C_ρ, the continuous function $|f|$ attains a maximum a. Let x, $y\in C$, then

$$z:=y+\frac{\rho}{\|y-x\|}(y-x)\in C_\rho$$

and

$$y=(1-\lambda)x+\lambda z\text{ with }\lambda=\frac{\|y-x\|}{\rho+\|y-x\|},$$

hence $f(y) \leq (1-\lambda)f(x) + \lambda f(y)$ yields

$$f(y) - f(x) \leq \lambda(f(z) - f(x)) \leq \tfrac{2a}{\rho} \|y - x\|.$$

Interchanging x and y we get $|f(y) - f(x)| \leq B\|y - x\|$ with $B = \frac{2a}{\rho}$ independent of x and y. Therefore f is Lipschitz-continuous on C. □

By Rademacher's Theorem (see Theorem 3.3.1), a Lipschitz-continuous function is differentiable $\mathcal{L}^n$-a.e. with Borel measurable gradient. Thus a convex function is differentiable $\mathcal{L}^n$-a.e. on the interior of its effective domain.

We will now show that the directional derivatives of a convex function exist everywhere.

Definition 1.2.3 (Directional derivatives)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$.

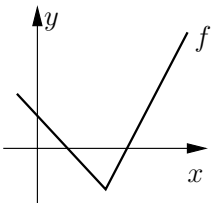
(i) If

$$\lim_{\lambda \to 0^+} \frac{f(x + \lambda y) - f(x)}{\lambda} \tag{1.7}$$

exists it is called the **one-sided directional derivative** of f in $x \in \mathbb{R}^n$ with respect to direction $y \in \mathbb{R}^n \backslash \{0\}$ and denoted by $f'(x; y)$.

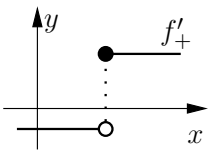
(ii) In the one dimensional case ($n = 1$), we distinguish between the **right derivative** f'_+ defined by

$$f'_+(x) := \lim_{\lambda \to 0^+} \frac{f(x + \lambda) - f(x)}{\lambda} = f'(x; 1)$$



and the **left derivative** $f'_-(x)$ defined by

$$f'_-(x) := \lim_{\lambda \to 0^+} \frac{f(x) - f(x - \lambda)}{\lambda} = -f'(x; -1).$$



Remark 1.2.4. Let $f : \mathbb{R} \to \mathbb{R}$ and $x \in \mathbb{R}$. If $f'_+(x) = f'_-(x)$, then f is differentiable at x with derivative $f'(x) = f'_+(x) = f'_-(x)$.

To prove the existence of the directional derivatives of a convex function we will first prove the 1-dimensional case.

Theorem 1.2.5 (Differentiability of convex functions in $\mathbb{R}$)**.** *Let $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be a convex function. Then f'_+ and f'_- exist on* $\operatorname{int}(\operatorname{dom} f)$ *and are non-decreasing functions. The inequality $f'_- \leq f'_+$ is valid, and with the exception of at most countably many points,*

$f'_- = f'_+$ holds (implying differentiability of f). Furthermore, f'_+ is continuous from the right and f'_- is continuous from the left and therefore, if f is differentiable on an open set, it is actually continuously differentiable.

Proof. The following proof is taken from [Sch93](Theorem 1.5.2). We will assume all arguments of f to be taken from $\operatorname{int}(\operatorname{dom} f)$. Let $0 < \lambda < \mu$. Then

$$f(x+\lambda) = f\left(\tfrac{\mu-\lambda}{\mu}x + \tfrac{\lambda}{\mu}(x+\mu)\right) \leq \tfrac{\mu-\lambda}{\mu}f(x) + \tfrac{\lambda}{\mu}f(x+\mu),$$

hence

$$\frac{f(x+\lambda)-f(x)}{\lambda} \leq \frac{f(x+\mu)-f(x)}{\mu}.$$

Analogously,

$$f(x-\lambda) = f\left(\tfrac{\mu-\lambda}{\mu}x + \tfrac{\lambda}{\mu}(x-\mu)\right) \leq \tfrac{\mu-\lambda}{\mu}f(x) + \tfrac{\lambda}{\mu}f(x-\mu),$$

hence

$$\frac{f(x)-f(x-\mu)}{\mu} \leq \frac{f(x)-f(x-\lambda)}{\lambda}.$$

For arbitrary λ, $\mu > 0$,

$$f(x) = f\left(\tfrac{\lambda}{\lambda+\mu}(x-\mu) + \tfrac{\mu}{\lambda+\mu}(x+\lambda)\right) \leq \tfrac{\lambda}{\lambda+\mu}f(x-\mu) + \tfrac{\mu}{\lambda+\mu}f(x+\lambda),$$

hence

$$\frac{f(x)-f(x-\mu)}{\mu} \leq \frac{f(x+\lambda)-f(x)}{\lambda}.$$

From the monotonicity and boundedness properties we just established, we can deduce the existence of f'_+ and f'_- as well as the inequality $f'_- \leq f'_+$. So, for $x < y$,

$$f'_-(x) \leq f'_+(x) \leq \frac{f(y)-f(x)}{y-x} \leq f'_-(y) \leq f'_+(y). \qquad (\star)$$

Hence f'_- and f'_+ are non-decreasing and thus have at most countably many discontinuities. At each continuity point x of f'_- the above inequality implies $f'_-(x) = f'_+(x)$ and hence the existence of the derivative $f'(x)$.

Let $x < y$. Using ($\star$) we get

$$\frac{f(y)-f(x)}{y-x} = \lim_{z\to x^+} \frac{f(y)-f(z)}{y-z} \geq \lim_{z\to x^+} f'_+(z),$$

hence $f'_+(x) \geq \lim_{z\to x^+} f'_+(z)$. Since f'_+ is non-decreasing, this implies that f'_+ is continuous from the right. Analogously one obtains that f'_- is continuous from the left. □

We will now prove that the directional derivatives of a convex function exist and are sub-linear functions of the direction.

Definition 1.2.6 (Sub-linear functions)**.** Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a function.

(i) f is called **positively homogeneous** if

$$f(\lambda x) = \lambda f(x), \qquad \forall \lambda > 0, x \in \mathbb{R}^n. \tag{1.8}$$

(ii) f is called **sub-additive** if

$$f(x+y) \leq f(x) + f(y), \qquad \forall x, y \in \mathbb{R}^n. \tag{1.9}$$

(iii) f is called **sub-linear** if it is positively homogeneous and sub-additive.

Remark 1.2.7. A proper sub-linear function f is convex, because

$$f((1-\lambda)x + \lambda y) \leq f(1-\lambda x) + f(\lambda y) = (1-\lambda) f(x) + \lambda f(y).$$

Theorem 1.2.8 (The directional derivative of a convex function is sub-linear)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function and $x \in \operatorname{int}(\operatorname{dom}(f))$. Then the directional derivatives at x exist and $f'(x;.) : \mathbb{R}^n \backslash \{0\} \to \mathbb{R}$ is sub-linear.*

Proof. Let $x \in \operatorname{int}(\operatorname{dom}(f))$ and fix a direction $y \in \mathbb{R}^n \backslash \{0\}$. Then $g : \mathbb{R} \to \mathbb{R}$ defined by $g(\lambda) := f(x + \lambda y)$ is convex and by Theorem 1.2.5 the right derivative of g in 0 exists. This yields

$$g'_+(0) = \lim_{\lambda \to 0^+} \frac{g(\lambda) - g(0)}{\lambda} = \lim_{\lambda \to 0^+} \frac{f(x+\lambda y) - f(x)}{\lambda} = f'(x; y).$$

So the directional derivatives of f in x exist.
Next we will prove the sub-linearity of $f'(x;.)$. Let $y \in \mathbb{R}^n \backslash \{0\}$ and λ, $\tau > 0$. We get

$$\begin{aligned} f'(x;\lambda y) &= \lim_{\tau \to 0^+} \frac{f(x+\tau\lambda y)-f(x)}{\tau} \\ &= \lim_{\tau \to 0^+} \lambda \frac{f(x+\tau\lambda y)-f(x)}{\tau\lambda} = \lambda f'(x;y), \end{aligned}$$

hence $f'(x;.)$ is positively homogeneous. For y, $z \in \mathbb{R}^n \backslash \{0\}$ convexity of f implies

$$f(x+\tau(y+z)) \leq f\left(\tfrac{1}{2}(x+2\tau y)+\tfrac{1}{2}(x+2\tau z)\right) \leq \tfrac{1}{2}f(x+2\tau y)+\tfrac{1}{2}f(x+2\tau z),$$

therefore

$$\begin{aligned} f'(x;y+z) &= \lim_{\tau \to 0^+} \frac{f(x+\tau(y+z))-f(x)}{\tau} \\ &\leq \lim_{\tau \to 0^+} \frac{f(x+2\tau y)-f(x)}{2\tau} + \frac{f(x+2\tau z)-f(x)}{2\tau} \leq f'(x;y)+f'(x;z). \end{aligned}$$

Hence $f'(x;.)$ is sub-additive and thus $f'(x;.)$ is sub-linear. □

We will show that the existence and sub-linearity of the directional derivatives imply, that a convex function is subdifferentiable on the interior of its domain.

Definition 1.2.9 (Subdifferentiability and subgradient)**.** Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper function and fix $x \in \mathrm{dom}(f)$. f is called **subdifferentiable** in x with **subgradient** $p \in \mathbb{R}^n$ if

$$f(y) \geq f(x) + \langle p, y-x \rangle + o(\|y-x\|) \text{ as } y \to x. \tag{1.10}$$

We call

$$\underline{\partial}_x f := \{p \in \mathbb{R}^n | p \text{ is subgradient of } f \text{ in } x.\} \tag{1.11}$$

the **subdifferential** of f in x.

Remark 1.2.10.

1. Analogously one defines superdifferentiability and supergradients. The superdifferential of f in x will be denoted by $\overline{\partial}_x f$.
2. If f is differentiable at x then obviously $\nabla f(x) \in \underline{\partial}_x f$.

3. If a function f has a supergradient as well as a subgradient in x, then f is differentiable at x and $\overline{\partial}_x f = \underline{\partial}_x f = \{\nabla f(x)\}$.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and subdifferentiable at x with subgradient p. Consider the affine function $h : \mathbb{R}^n \to \mathbb{R}$ defined by $h(z) = f(x) + \langle p, z - x\rangle$, then $f(x) = h(x)$. So locally around x we have $h \leq f$. This even holds globally, because suppose there is $y_0 \in \mathbb{R}^n$ such that $f(y_0) < h(y_0)$. Let $\lambda \in (0,1)$, then

$$\begin{aligned} f(\lambda y_0 + (1-\lambda)x) &\leq \lambda f(y_0) + (1-\lambda) f(x) \\ &< \lambda h(y_0) + (1-\lambda) h(x) = h(\lambda y_0 + (1-\lambda)x). \end{aligned}$$

Therefore

$$f(x + \epsilon(y_0 - x)) < h(x + \epsilon(y_0 - x))$$

for all $\epsilon > 0$ in contradiction to $h \leq f$ locally around x. Thus $h \leq f$ holds globally on $\mathbb{R}^n$ and therefore a convex function $f : \mathbb{R}^n \to \mathbb{R}$ is subdifferentiable at x with subgradient p if and only if

$$f(z) \geq f(x) + \langle p, z - x\rangle, \forall z \in \mathbb{R}^n. \tag{1.12}$$

Also, note that

$$\operatorname{epi}(h) = \overline{H}^{+}_{(p,-1),\langle x,p\rangle - f(x)}.$$

and $h \leq f$ if and only if

$$\operatorname{epi}(h) \supseteq \operatorname{epi}(f).$$

Thus $H_{(p,-1),\langle x,p\rangle - f(x)}$ is a **supporting hyperplane** of $\operatorname{epi}(f)$.

Theorem 1.2.11 (Subdifferentiability of convex functions)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and $x \in \operatorname{int}(\operatorname{dom}(f))$. $p \in \mathbb{R}^n$ is subgradient of f in x if and only if*

$$f'(x; y) \geq \langle p, y\rangle \qquad \forall y \in \mathbb{R}^n. \tag{1.13}$$

Thus

$$\underline{\partial}_x f = \{p \in \mathbb{R}^n | \forall y \in \mathbb{R}^n : \langle p, y\rangle \leq f'(x;y)\} \tag{1.14}$$

and $\underline{\partial}_x f$ is a compact convex set.

Proof. Let $x \in \operatorname{int}(\operatorname{dom}(f))$ and $y \in \mathbb{R}^n \backslash \{0\}$ arbitrary. We recall from the proof of Theorem 1.2.5 that the quotient $\frac{f(x+\lambda y)-f(x)}{\lambda}$ is a decreasing function as $\lambda \to 0^+$, thus

$$\frac{f(x+\lambda y)-f(x)}{\lambda} \geq f'(x;y) \geq \langle p, y\rangle, \qquad \forall \lambda > 0.$$

Analogously we get

$$\frac{f(x-\lambda y)-f(x)}{\lambda} \geq \langle p, -y\rangle, \qquad \forall \lambda > 0.$$

Combining these inequalities we get

$$f(x+\lambda y)-f(x) \geq \langle p, \lambda y\rangle, \qquad \forall \lambda \in \mathbb{R}.$$

Thus, by putting $z := x + \lambda y$, we get

$$f(z) \geq f(x) + \langle p, z-x\rangle, \qquad \forall z \in \mathbb{R}^n.$$

Therefore p satisfies condition (1.13) if and only if p is subgradient of f in x.
Let $x \in \operatorname{int}(\operatorname{dom}(f))$. We will show that $\underline{\partial}_x f$ is non-empty. $(x, f(x))$ is a boundary point of $\operatorname{epi}(f)$. Since $\{(x, f(x))\}$ is compact and

$$\operatorname{int}(\operatorname{epi}(f)) = \{(x,\mu) | f(x) < \mu\}$$

is non-empty and open, we can use the Hahn-Banach separation Theorem and get $u \in \mathbb{R}^{n+1}$ and $\alpha > 0$ such that

$$H^+_{u,\alpha} \supseteq \operatorname{int}(\operatorname{epi}(f))$$

and $(x, f(x)) \in H_{u,\alpha}$, i.e. $\langle u, (x, f(x))\rangle = \alpha$. Let $u = (p, -v)$ for $p \in \mathbb{R}^n$ and $v \in \mathbb{R}$. Since

$(x,\mu) \in \text{int}(\text{epi}(f))$ for all $\mu > f(x)$, we have

$$\langle p, x\rangle - vf(x) = \alpha > \langle u, (x,\mu)\rangle = \langle p, x\rangle - v\mu,$$

hence $v > 0$. Put $q := \frac{p}{v}$, $\beta := \frac{\alpha}{v}$ and define the affine function $h : \mathbb{R}^n \to \mathbb{R}$ by

$$h(z) := \langle q, z\rangle - \beta.$$

Then $h(x) = \frac{1}{v}(\langle p, x\rangle - \alpha) = f(x)$ and $h \leq f$, because

$$\text{epi}(h) = \overline{H}^+_{(q,-1),\beta} = \overline{H}^+_{(p,v),\alpha} \supseteq \text{epi}(f).$$

Therefore $q \in \underline{\partial}_x f$.
Equality (1.14) is obvious from condition (1.13). Thus, we can write

$$\underline{\partial}_x f = \bigcap_{y\in\mathbb{R}^n} \{p \in \mathbb{R}^n | \langle p, y\rangle \leq f'(x;y)\} = \bigcap_{y\in\mathbb{R}^n} \overline{H}^+_{y,f'(x;y)},$$

hence $\underline{\partial}_x f$ is convex and closed as an intersection of convex and closed sets (closed half-spaces) by Proposition 1.1.3.
To show that $\underline{\partial}_x f$ is compact we need to show that $\underline{\partial}_x f$ is bounded. We notice that

$$\sup_{p\in\underline{\partial}_x f} \langle p, y\rangle \leq f'(x;y) < +\infty \qquad \forall y \in \mathbb{R}^n,$$

hence $\underline{\partial}_x f$ has to be bounded and therefore $\underline{\partial}_x f$ is compact. □

Remark 1.2.12.

1. Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function. The subdifferential of f in $x \in \text{int}(\text{dom}(f))$ is given by the closed interval $\underline{\partial}_x f = [f'_-(x), f'_+(x)]$.

2. For a convex set $C \subseteq \mathbb{R}^n$ we call the sub-linear function defined by

   $$h(y|C) := \sup\{\langle x, y\rangle | x \in C\}$$

   the **support function** of C.
   It is a basic result from convex analysis (see Corollary 13.2.1 in [Roc70]) that every sub-linear function f is the support function of a closed convex set C, namely $C = \{y | \langle x, y\rangle \leq f(x)\}$. So the previous Theorem actually states that the directional

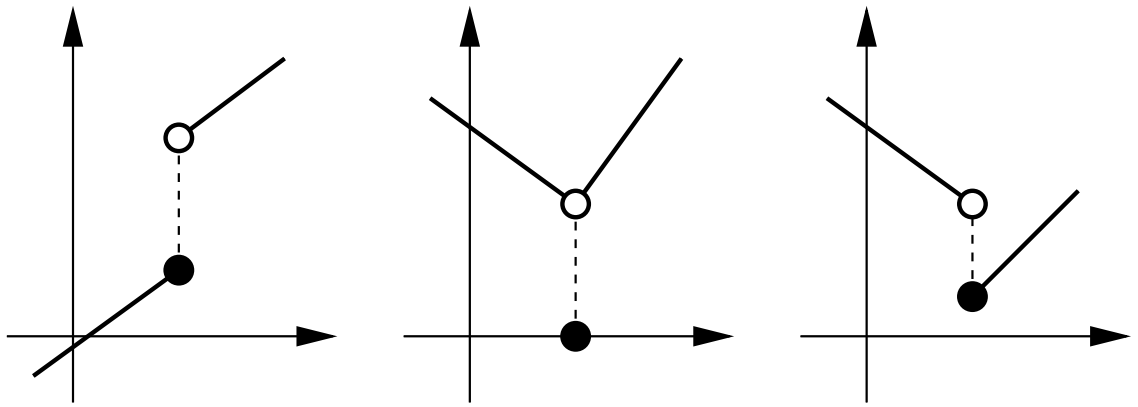


Figure 1.2: Examples of lower semi-continuous functions.

derivative is the support function of the subdifferential of f, i.e.

$$f'(x;y) = h(y|\underline{\partial}_x f).$$

This result can also be found in [Roc70] (Theorem 23.4).

3. If the subdifferential of f in x is single-valued, i.e. $\underline{\partial}_x(f) = \{p\}$, we have

$$f'(x;y) = h(y|\underline{\partial}_x f) = \langle p, y\rangle .$$

So the directional derivative is a linear function and therefore f is differentiable at x with $\nabla f(x) = p$.

1.3 Conjugates of Convex Functions

By Theorem 1.1.5 a closed convex set is the intersection of all closed half-spaces containing it. We will show that a lower semi-continuous (see the definition below) convex function is the point-wise supremum of all affine functions less or equal to it.

Definition 1.3.1. (Lower semi-continuous (lsc) functions)
A function $f : A \subseteq \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ is called **lower semi-continuous (lsc)** in $x \in A$, if and only if

$$f(x) \leq \liminf_{y\to x} f(y) := \lim_{n\to\infty} \inf\{f(y)|y \in B_{1/n}(x)\backslash\{x\}\}.$$

Remark 1.3.2.

1. A function $f : \mathbb{R}^n \to \mathbb{R}$ is lsc on $\mathbb{R}^n$ if and only if the epigraph of f is closed. This can be seen easily: f is lsc in x if and only if for any sequence $(\mu_i)_{i\in\mathbb{N}}$ in $\mathbb{R}$ and $(x_i)_{i\in\mathbb{R}^n}$ with $\mu_i \to \mu$, $x_i \to x$ and $\mu_i \geq f(x_i)$ we have $\mu \geq f(x)$. But this is just the closure of $\text{epi}(f) = \{(x,\mu)\mathbb{R}^{n+1} | f(x) \leq \mu\}$.
   In view of this relation a lsc function is sometimes called **closed**.

2. Given a proper function f, one defines the **lower semi-continuous hull** of f by

$$\text{cl}f = \sup\{g | g \text{ is lsc and } g \leq f\}.$$

   Then

$$\text{epi}(\text{cl}f) = \bigcup\{\text{epi}(g) | g \text{ is lsc and } g \leq f\}$$

   is closed and $\text{cl}f(x) \leq \liminf_{y\to x} f(y)$.

3. Let f be a convex function. Then $\text{cl}f$ is convex and since a convex function is continuous in $\text{int}(\text{dom}(f))$ we have $\text{cl}f = f$ there. Furthermore, because $f = +\infty$ outside $\text{dom}(f)$ we actually have $\{\text{cl}f \neq f\} \subseteq \partial\text{dom}(f)$, where $\partial\text{dom}(f)$ denotes the boundary of $\text{dom}(f)$.

Theorem 1.3.3. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be lsc and convex. Then f is the point-wise supremum of all affine functions h such that $h \leq f$, i.e.*

$$f(x) = \sup\{h(x) | h \text{ is affine and } h \leq f\}.$$

Idea of the proof. This can be deduced from the following: f is lsc and convex if and only if $\text{epi}(f)$ is closed and convex. A closed convex set is the intersection of all closed half-spaces containing it (Theorem 1.1.5). Now, one can show that the intersection of half-spaces of the form $\overline{H}^+_{(u,-1),\alpha}$ with $u \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$ and $\text{epi}(f) \subseteq \overline{H}^+_{(u,-1),\alpha}$ are sufficient to generate $\text{epi}(f)$. This leads to the above result, because $\text{epi}(h) = H_{(u,-1),\alpha}$. For a rigorous proof of this fact we refer to [Roc70] Theorem 12.1. □

Definition 1.3.4 (Conjugate function)**.** Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper function. The function $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ defined by

$$f^*(y) := \sup_{x \in \mathbb{R}^n} \langle x, y \rangle - f(x) \tag{1.15}$$

is called the **conjugate** of f.

Remark 1.3.5.

1. The conjugate function is also sometimes called the **Legendre-Fenchel transform**.

2. The conjugate of a proper function which is bounded from below by an affine function is a lower semi-continuous convex function. This can easily be seen: Let f be a proper function and h affine such that $h \leq f$. Since f is proper we have $\{f = +\infty\} \neq \mathbb{R}^n$ and therefore $\{f^* = -\infty\} = \emptyset$. Since h is affine, there are $u \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$, such that $h(x) = \langle x, u \rangle - \alpha$. We get

   $$\langle x, u \rangle - f(x) \leq \alpha < +\infty \qquad \forall x \in \mathbb{R}^n$$

   hence $f^*(u) \leq \alpha$ and thus $\{f^* = +\infty\} \neq \mathbb{R}^n$. So f^* is a proper function. To show lower semi-continuity and convexity we rewrite the epigraph of f^* as

   $$\operatorname{epi}(f^*) = \bigcap_{x \in \mathbb{R}^n} \{(y, \mu) \in \mathbb{R}^{n+1} | \mu \geq \langle x, y \rangle - f(x)\} = \bigcap_{x \in \mathbb{R}^n} \overline{H}^+_{(x,-1),f(x)}.$$

   Therefore $\operatorname{epi}(f^*)$ is a closed convex set as an intersection of closed half-spaces and non-empty since f^* is proper. Thus f^* is lower semi-continuous and convex.

3. Definition (1.15) implies

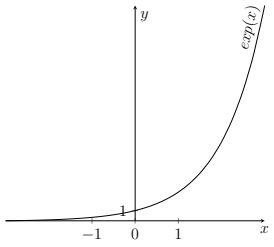
   $$\langle x, y \rangle \leq f(x) + f^*(y) \qquad \text{for all } x, y \in \mathbb{R}^n. \tag{1.16}$$

   This inequality is known as **Fenchel's inequality**. Its equality cases will be very important to us.

Example 1.3.6.

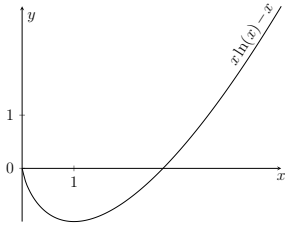
(a) Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \exp(x).$$



f is convex and its conjugate $f^* : \mathbb{R} \to \mathbb{R}$ is given by

$$f^*(y) = \begin{cases} y\ln(y) - y & \text{if } y > 0, \\ 0 & \text{if } y = 0, \\ +\infty & \text{else.} \end{cases}$$



(b) Fix $p \in (1, +\infty)$ and define $f : \mathbb{R}^n \to \mathbb{R}$ by

$$f(x) = \frac{\|x\|^p}{p}.$$

f is convex and its conjugate $f^* : \mathbb{R}^n \to \mathbb{R}$ is given by

$$f^*(y) = \frac{\|y\|^q}{q},$$

where $q \in (1, +\infty)$ such that $\frac{1}{p} + \frac{1}{q} = 1$.
For $p = q = \frac{1}{2}$ we have $f = \frac{\|x\|^2}{2} = f^*$. One can show that this is the only solution for a convex function f such that $f = f^*$.

We will now give a characterization of the subdifferential and state a duality correspondence of lower semi-continuous convex functions. The following results are mostly taken from [Vil03].

Proposition 1.3.7 (Characterization of the subdifferential)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower semi-continuous convex function. Then*

$$y \in \underline{\partial}_x f \Leftrightarrow \langle x, y \rangle = f(x) + f^*(y). \tag{1.17}$$

Proof. By Fenchel's inequality (1.16) we have

$$\begin{aligned}\langle x,y\rangle = f(x)+f^*(y) &\Leftrightarrow \langle x,y\rangle \geq f(x)+f^*(y)\\ &\Leftrightarrow \langle x,y\rangle \geq f(x)+\langle y,z\rangle - f(z) \qquad \forall z\in\mathbb{R}^n\\ &\Leftrightarrow f(z)\geq f(x)+\langle y,z-x\rangle \qquad \forall z\in\mathbb{R}^n\\ &\Leftrightarrow y\in \underline{\partial}_x f.\end{aligned}$$

□

Theorem 1.3.8 (Duality correspondence of convex functions)**.** *Let $f:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ be a proper function. Then f is a lower semi-continuous convex function if and only if $f^{**}=f$.*

Proof. From Remark 1.3.5(2) it is clear that $f^{**}=f$ implies that f is lower semi-continuous and convex. So all we need to show is that a lower semi-continuous convex function satisfies $f^{**}=f$. Using Fenchel's inequality (1.16) we get

$$f(x)\geq \sup_{y\in\mathbb{R}^n}\langle x,y\rangle - f^*(y) = f^{**}(x),$$

hence $f\geq f^{**}(x)$.
On the other hand, for $x\in\operatorname{int}(\operatorname{dom}(f))$, by Theorem 1.2.11, we have $\underline{\partial}_x f\neq\emptyset$, thus we can choose $y\in\underline{\partial}_x f$. Using Proposition 1.3.7 we have $f(x)+f^*(y)=\langle x,y\rangle$ and therefore

$$f(x)\leq \sup_{y\in\mathbb{R}^n}\langle x,y\rangle - f^*(y) = f^{**}(x),$$

which implies $f=f^{**}$ on $\operatorname{int}(\operatorname{dom}(f))$. So, if $\operatorname{dom}(f)=\mathbb{R}^n$, then $f=f^{**}$.
To finish the proof we will use the same strategy as in the proof of Proposition 2.5 in [Vil03]. We will regularize f by infimal convolution (see Proposition 1.1.10(iv)): Let $g_\epsilon(x):=\frac{\|x\|^2}{2\epsilon}$ and

$$f_\epsilon(x):=f\square g_\epsilon(x)=\inf_{y\in\mathbb{R}^n} f(x+y)+g_\epsilon(y).$$

We will show that $\lim_{\epsilon\to 0^+} f_\epsilon = f$. First, notice that for $y=0$ we get $f_\epsilon(x)\leq f(x)$ and therefore $\lim_{\epsilon\to 0^+} f_\epsilon\leq f$. To show the other inequality, we fix an arbitrary affine function $h(x)=\langle x,u\rangle-\alpha$ with $h\leq f$. We get

$$f_\epsilon(x+y)+g(y)\geq\langle x+y,u\rangle-\alpha+\frac{\|y\|^2}{2\epsilon}.$$

The right-hand side of this inequality is a quadratic function in y and therefore attains an unique minimum at $y = -\epsilon u$. Thus

$$\begin{aligned} f_\epsilon(x) &\geq \langle x - \epsilon u, u\rangle - \alpha + \frac{\|\epsilon u\|^2}{2\epsilon} \\ &= \langle x, u\rangle - \alpha - \frac{\epsilon \|u\|^2}{2} = h(x) - \frac{\epsilon \|u\|^2}{2}. \end{aligned}$$

Thus $\lim_{\epsilon \to 0^+} f_\epsilon \geq h$ for all affine h with $h \leq f$. Since f is lower semi-continuous and convex, we can use Theorem 1.3.3 to get

$$f = \sup_{\substack{h \leq f \\ h \text{ affine}}} h \leq \lim_{\epsilon \to 0^+} f_\epsilon,$$

hence $\lim_{\epsilon \to 0^+} f_\epsilon = f$.
Since $\operatorname{dom}(f_\epsilon) = \mathbb{R}^n$, we have $f_\epsilon^{**} = f_\epsilon$. Furthermore, $f_\epsilon \leq f$ implies $f_\epsilon^* \geq f^*$ which in turn implies $f_\epsilon^{**} \leq f^{**}$. We conclude

$$f^{**}(x) \geq \lim_{\epsilon \to 0^+} f_\epsilon^{**}(x) = \lim_{\epsilon \to 0^+} f_\epsilon(x) = f(x),$$

thus $f = f^{**}$. □

Remark 1.3.9.

1. Let f be a lower semi-continuous convex function. Using Theorem 1.3.8 and Proposition 1.3.7 we have

$$y \in \underline{\partial}_x f \Leftrightarrow \langle x, y\rangle = f(x) + f^*(y) = f^{**}(x) + f^*(y) \Leftrightarrow x \in \underline{\partial}_y f^*.$$

2. Let f be lower semi-continuous and strictly convex. For $x_1 \neq x_2 \in \operatorname{int}(\operatorname{dom}(f))$ one can show that

$$\underline{\partial}_{x_1} f \cap \underline{\partial}_{x_2} f = \emptyset.$$

Therefore $y \in \underline{\partial}_x f$ implies that $\underline{\partial}_y f^* = \{x\}$, hence f^* is differentiable at y with $\nabla f^*(y) = x$. If f is also differentiable at x, then one has $\underline{\partial}_x f = \{\nabla f(x)\}$ and thus

$$\nabla f^*(\nabla f(x)) = x.$$

Since $y \in \underline{\partial}_x f \Leftrightarrow x \in \underline{\partial}_y f^*$ (see Proposition 1.3.7), we also have

$$\nabla f(\nabla f^*(y)) = y.$$

A convex function is differentiable $\mathcal{L}^n$-a.e. in the interior of its effective domain with Borel measurable gradient (see Theorem 1.2.2 and Theorem 3.3.1). So we have $\nabla f^* \circ \nabla f(x) = x$ $\mathcal{L}^n$-a.e. on $\operatorname{int}(\operatorname{dom}(f))$. Hence we may use ∇f^* as inverse to ∇f.

Chapter 2

Introduction to Optimal Transport

In this chapter we will first motivate and define the **Monge Problem (MP)** and a relaxation of it, the **Monge-Kantorovich Problem (MKP)**. We will then prove a dual representation of the MKP, called the Kantorovich Duality. In the proof we will introduce c-transform and c-concavity, which can be seen as generalization of conjugates and convexity.

In the final part of this chapter, we will solve the MKP explicitly on the real line.

2.1 Motivation and Definitions

Before we give a formal definition of the Optimal Transport Problem let us consider the following economic example, which was given by Cédric Villani in [Vil09]:

> "Consider a large number of bakeries, producing loaves, that should be transported each morning to cafés where consumers will eat them. The amount of bread that can be produced at each bakery, and the amount that will be consumed at each café are known in advance, and can be modeled as probability measures (there is a "density of production" and a "density of consumption") on a certain space, which in our case would be Paris (equipped with the natural metric such that the distance between two points is the length of the shortest path joining them). The problem is to find in practice where each unit of bread should go, in such a way as to minimize the total transport cost."

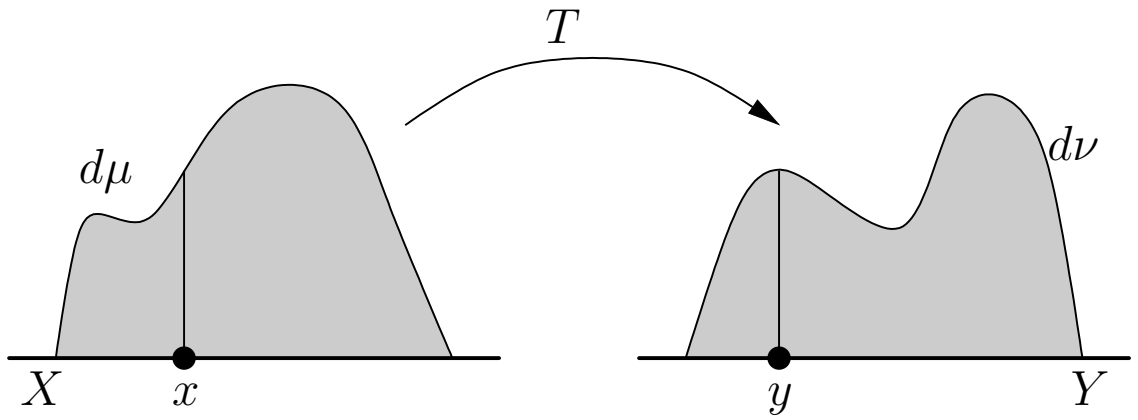
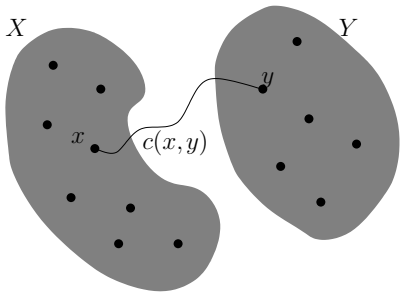


Figure 2.1: In the Optimal Transport Problem we are looking for transport plans $T : X \to Y$ that transport mass from X distributed by μ to Y according to ν, i.e. $T\#\mu = \nu$.

To put this in mathematical terms, the "space of bakeries" will be called X and the "space of cafés" Y. The "density of production" is then a probability measure μ on X and the "density of consumption" is a probability measure ν on Y.

In order to move loaves of bread from a bakery $x \in X$ to a café $y \in Y$ some kind of effort has to be made. This will be measured by a measurable cost function $c : X \times Y \to [0, +\infty]$. So the cost of moving x to y equals $c(x, y)$. For example, $c(x, y)$ could be the length of the shortest path between x and y.



The problem can now be described as finding measurable maps $T : X \to Y$ which assign to each bakery x a café $y = T(x)$ such that the "density of production" μ matches the "density of consumption" ν. This condition is characterized by the following

$$T\#\mu = \nu, \tag{2.1}$$

by which we mean that for every measurable set $B \subseteq Y$ we have $\mu[T^{-1}(B)] = \nu[B]$, where $T^{-1}(B) = \{x \in X | T(x) \in B\}$ is the preimage of B under T. $T\#\mu$ is called the **push-forward** of μ by T. Maps T which satisfy (2.1) will be called **transport plans**.

To get a better understanding of condition (2.1) we may consider μ and ν to be discrete. Then $\mu(x)$ is the quantity of bread produced at bakery x and $\nu(y)$ is the quantity of bread consumed at café y. Further, a map $T : X \to Y$ is a valid transport plan if and

only if

$$\nu(y) = \sum_{x:\, y=T(x)} \mu(x),$$

i.e. the quantity of bread moved to a café y by T has to match the quantity of bread which is consumed there.

Given a transport plan T we will assign a total cost $I_c(T)$ to it by letting

$$I_c(T) := \int_X c(x, T(x)) d\mu(x). \tag{2.2}$$

We can now formulate the Optimal Transport Problem.

Definition 2.1.1 (Monge Problem, MP)**.** Let μ and ν be probability measures on probability spaces X respectively Y and let $c : X \times Y \to [0, +\infty]$ be a measurable cost function. The **Monge Problem (MP)** is to find a transport plan $T^* : X \to Y$ such that

$$I_c(T^*) = \inf_{T \in \mathcal{T}(\mu,\nu)} I_c(T),$$

where

$$\mathcal{T}(\mu,\nu) := \{T : X \to Y |\ T \text{ is a transport plan between } \mu \text{ and } \nu.\}.$$

The MP may not have a solution, i.e. the set of transport plans $\mathcal{T}(\mu,\nu)$ may be empty or there may not exist a minimizer T^*.

Example 2.1.2 (Solvability of MP)**.**

(a) Let $X = \{x\}$ be a space with only a single point and $Y = \{y_1, y_2\}$ a space with two points. Let μ be the trivial probability measure on X and ν the probability measure that splits its mass evenly on the two points of Y, i.e. $\nu[\{y_1\}] = \nu[\{y_2\}] = \frac{1}{2}$. Clearly, there are only two maps to consider: $T_1(x) = y_1$ or $T_2(x) = y_2$. But neither of those maps are transport plans, since the push-forward of μ by neither T_1 nor T_2 is equal to ν. In fact $\mathcal{T}(\mu,\nu) = \emptyset$.

(b) Let

$$X = [0,1] \times \{0\} \text{ and } Y = [0,1] \times \{\pm 1\}.$$

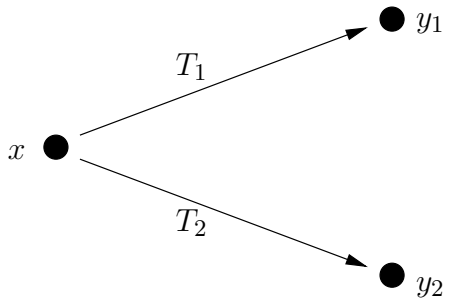


Figure 2.2: Sketch of Example 2.1.2(a). A situation where no transport plans exist.

Let μ be uniformly distributed on $[0,1]\times\{0\}$, i.e.

$$\mu(A\times\{0\}) = \mathcal{L}^1(A)$$

where $\mathcal{L}^1$ is the Lebesgue measure on $\mathbb{R}$. Let ν be uniformly distributed on Y, i.e.

$$\nu((B_1\times\{-1\})\times(B_2\times\{+1\})) = \tfrac{1}{2}\mathcal{L}^1(B_1)+\tfrac{1}{2}\mathcal{L}^1(B_2).$$

One can think of X as one line and Y as two parallel lines, one to the left and one to the right of X. The cost function shall be

$$c((x,0),(y,z)) = \|(x,0)-(y,z)\|^2 = (x-y)^2+1$$

for all x, $y\in[0,1]$ and $z\in\{\pm1\}$. Then one can define a sequence of transport plans $(T_i)_{i\in\mathbb{N}}$ in the following way: First define T_1 by

$$T_1((x,0)) = \begin{cases} (2x,-1), & x\in[0,\frac{1}{2}),\\ (2x-1,+1), & x\in[\frac{1}{2},1]. \end{cases}$$

So T_1 basically cuts the line X in half and assigning the lower half to the left side of Y and the upper half to the right side. It is easy to verify that T_1 is a valid transport plan from μ to ν. The total cost $I_c(T_1)$ is given by

$$I_c(T_1) = \int_0^{\frac{1}{2}} 1+x^2dx + \int_{\frac{1}{2}}^1 1+(1-x)^2dx = 2\int_0^{\frac{1}{2}} 1+x^2dx = \frac{13}{12}$$

Next we define T_2 by cutting X into four equal parts and assign two to the left of

Y and two to the right, i.e.

$$T_2((x,0)) = \begin{cases} (-1, 2x), & x \in [0, \frac{1}{4}), \\ (+1, 2x - \frac{1}{2}), & x \in [\frac{1}{4}, \frac{1}{2}), \\ (-1, 2x - \frac{1}{2}), & x \in [\frac{1}{2}, \frac{3}{4}), \\ (+1, 2x - 1), & x \in [\frac{3}{4}, 1]. \end{cases}$$

This again yields a valid transport plan and the total cost is given by

$$I_c(T_2) = 4 \int_0^{\frac{1}{4}} 1 + x^2 dx = \frac{49}{48}.$$

Proceeding in this manner, T_i will be the transport plan that cuts X into 2^i equal parts and distributes them to the left and right of Y. The total cost $I_c(T_i)$ is given by

$$I_c(T_i) = 2^i \int_0^{2^{-i}} 1 + x^2 dx = 1 + \frac{1}{3 \cdot 4^i}.$$

One can show that

$$\inf_{T \in \mathcal{T}(\mu,\nu)} I_c(T) = \lim_{i \to \infty} I_c(T_i) = 1,$$

thus the optimal transport cost of the MP is approximated by the sequence $(T_i)_{i \in \mathbb{N}}$. But the sequence does not converge to a map T^*. In fact, there is no transport plan $T^* \in \mathcal{T}(\mu, \nu)$ such that $I_c(T^*) = 1$, therefore the MP has no solution.

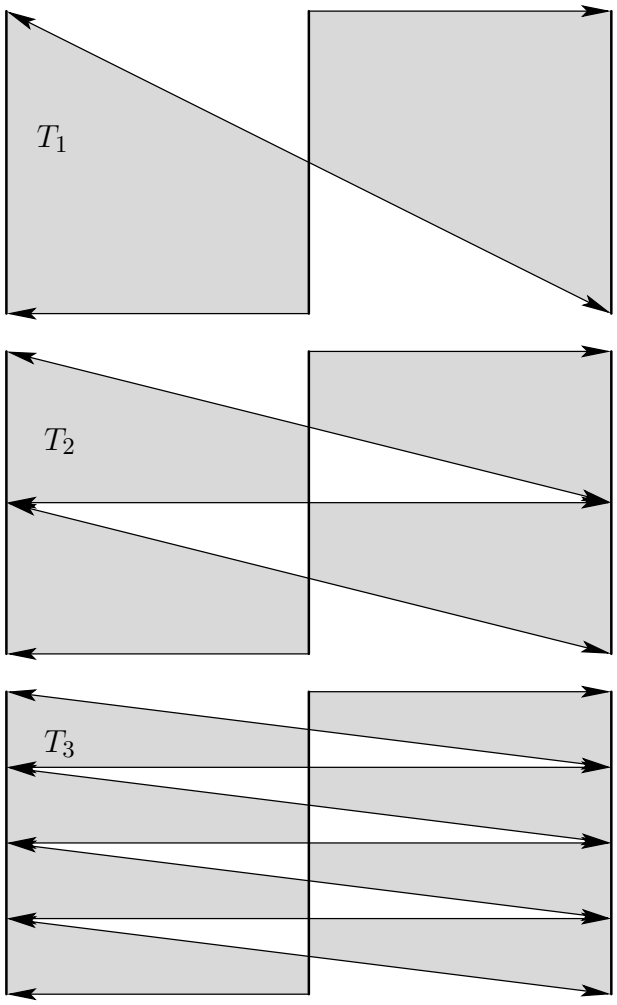


Figure 2.3: Sketches of the transport plans T_1, T_2 and T_3 as defined in Example 2.1.2(b). This sequence approximates the optimal transport cost, but does not converge to a map.

Example 2.1.3 (Nonlinear constraint on T)**.** Let $X = Y = \mathbb{R}^n$ and μ, ν be probability measures on $\mathbb{R}^n$ which are absolutely continuous with respect to the Lebesgue measure $\mathcal{L}^n$, i.e.

$$d\mu(x) = f(x)dx \text{ and } d\nu(y) = g(y)dy.$$

Now condition (2.1) for a transport plan T can be formulated as

$$\int\limits_{T^{-1}(B)} f(x)dx = \int\limits_{B} g(y)dy,$$

whenever $B \subseteq R^n$ is measurable. If we assume T to be a $\mathcal{C}^1$-diffeomorphism from $\mathbb{R}^n$ onto itself we can use the **change of variable formula** to get

$$\int\limits_{T^{-1}(B)} f(x)dx = \int\limits_{T^{-1}(B)} g(T(x))|\det dT(x)|dx.$$

So if μ and ν are absolutely continuous to $\mathcal{L}^n$ with densities f respectively g, then a differentiable function $T : \mathbb{R}^n \to \mathbb{R}^n$ is a transport plan if and only if

$$f(x) = g(T(x))|\det dT(x)|, \mu\text{-a.e.}$$

This condition is obviously nonlinear.

The MP was relaxed by Kantorovich in the forties[1] [Kan42, Kan48]. Instead of looking for transport plans, Kantorovich considered probability measures γ on the product space $X \times Y$ with marginals μ and ν, i.e. for all measurable $A \subseteq X$ and measurable $B \subseteq Y$:

$$\gamma(A \times Y) = \mu(A) \text{ and } \gamma(X \times B) = \nu(B). \tag{2.3}$$

A measure γ satisfying (2.3) will be called a **transference plan**.

For a transference plan γ one can interpret the quantity $\gamma(A \times B)$ as the amount of mass that is moved from $A \subseteq X$ to $B \subseteq Y$. Therefore the condition $\gamma(A \times Y) = \mu(A)$ means that all the mass $\mu(A)$ in A is moved somewhere to Y and the condition $\gamma(X \times B) = \nu(B)$ means that the amount of mass that is moved to B is $\nu(B)$.

[1] Leonid Kantorovich was awarded a Nobel prize in economics in 1975 for related work.

Every transport plan T induces an unique transference plan $\gamma_T := (\mathrm{id} \times T)\#\mu$. Thus transference plans can be seen as relaxations of transport plans.

The total cost of a transference plan γ is defined by

$$I_c(\gamma) := \int_{X \times Y} c(x,y) d\gamma(x,y).$$

We can easily verify that $I_c(T) = I_c(\gamma_T)$. The set of transference plans is denoted by

$$\Pi(\mu,\nu) := \{\gamma \in \mathcal{P}(X \times Y) |\ \gamma \text{ is a transference plan between } \mu \text{ and } \nu.\}.$$

Note that $\Pi(\mu,\nu) \neq \emptyset$ since $\mu \times \nu \in \Pi(\mu,\nu)$. Therefore the optimal transport cost I_c defined by

$$I_c := \inf_{\gamma \in \Pi(\mu,\nu)} I_c(\gamma)$$

always exists (but may be $+\infty$).

We can now reformulate the Optimal Transport Problem.

Definition 2.1.4 (Monge-Kantorovich Problem, MKP)**.** Let μ and ν be probability measures on probability spaces X respectively Y and let $c : X \times Y \to [0, +\infty]$ be a measurable cost function. The **Monge-Kantorovich Problem (MKP)** is to find a transference plan $\gamma^* \in \Pi(\mu,\nu)$ such that

$$I_c(\gamma^*) = I_c.$$

The MKP always admits solutions under quite general assumptions on the spaces X, Y and the cost function c.

Theorem 2.1.5 (Solvability of the Monge-Kantorovich Problem)**.** *Let X, Y be Polish probability spaces (i.e. complete, metric and separable) and μ, ν Borel probability measures on X respectively Y. Let $c : X \times Y \to [0, +\infty]$ be a lsc cost function. Then there exists an optimal transference plan $\gamma^* \in \Pi(\mu,\nu)$.*

Remark 2.1.6. Analogously to Definition 1.3.1, a function $f : X \to \mathbb{R} \cup \{\pm\infty\}$ on a Polish space X is lsc if f is lsc in x, i.e.

$$f(x) \leq \liminf_{y \to x} f(y) := \lim_{n \to \infty} \inf\{f(y) | y \in B_{1/n}(x) \backslash \{x\}\},$$

for all $x \in X$.

Before we prove Theorem 2.1.5 we need to recall some results from measure theory on Polish spaces.

Definition 2.1.7 (Borel measure, regular measures, tightness, weak convergence)**.** Let X be a Polish probability space.

(i) A measure μ is called a **Borel measure** if it is defined on the Borel σ-algebra on X, i.e. the smallest σ-algebra on X that contains all open subsets of X. The set of Borel probability measures on X is denoted by $\mathcal{P}(X)$.

(ii) A probability measure μ on X is called **regular** if for all measurable $A \subseteq X$ and $\epsilon > 0$ there is a closed set F and an open set G such that $F \subseteq A \subseteq G$ and $\mu(G \backslash F) \leq \epsilon$. Equivalently, A is regular if and only if

$$\mu(A) = \sup\{\mu(F) | F \subseteq A,\ F \text{ closed}\}$$

and

$$\mu(A) = \inf\{\mu(G) | G \subseteq A,\ G \text{ open}\}.$$

(iii) A subset $B \subseteq \mathcal{P}(X)$ of probability measures on X is called **tight** if for all $\epsilon > 0$ there exists a compact set K_ϵ such that

$$\sup_{\mu \in B} \mu(X \backslash K_\epsilon) \leq \epsilon.$$

Further, we call a measure $\mu \in \mathcal{P}(X)$ tight if $\{\mu\}$ is tight.

(iv) A sequence $(\mu_n)_{n \in \mathbb{N}}$ of probability measures on $\mathcal{P}(X)$ is said to **converge weakly** to a probability measure $\mu \in \mathcal{P}(X)$, $\mu_n \xrightarrow{\text{w}} \mu$, if and only if

$$\lim_{n \to \infty} \int_X f d\mu_n = \int_X f d\mu$$

for all bounded continuous functions $f : X \to \mathbb{R}$. (Equivalently one can choose bounded Lipschitz-continuous or lsc and bounded from below functions.)

Remark 2.1.8.

1. Any Borel probability measure on a Polish probability space is regular.
2. Any probability measure on a Polish probability space is tight. Furthermore, for a tight sequence of probability measures, there exists a measure $\mu \in \mathcal{P}(X)$ and a subsequence $(\mu_n)_{n\in\mathbb{N}}$, such that $\mu_n \xrightarrow{\text{w}} \mu$, this is **Prokhorov's Theorem**.

Remark 2.1.9.

1. Any non-negative lsc function $f : X \to [0, +\infty]$ on a metric space can be written as the supremum of a non-decreasing sequence of bounded, uniformly continuous and non-negative functions. One can choose

$$f_n(x) := \inf_{y\in X} \left[\min(f(y), n) + nd(x, y)\right],$$

then $0 \leq f_n \leq n$ and one easily shows that

$$f_n(x) - f_n(y) \leq nd(x, y),$$

thus f_n is bounded and uniformly continuous.
Furthermore, for $m \leq n$, we have

$$f_m(x) \leq \min(f(y), m) + md(x, y) \leq \min(f(y), n) + nd(x, y) \qquad \forall y \in X,$$

thus $f_m(x) \leq f_n(x)$ and therefore $(f_n)_{n\in\mathbb{N}}$ is a non-decreasing sequence.
Finally, we have $f_n(x) \leq \min(f(x), n)$, hence $f_n(x) \leq f(x)$. On the other hand, for all $n \in \mathbb{N}$ there exists y_n such that

$$f_n(x) + \tfrac{1}{n} \geq \min(f(y_n), n) + nd(x, y_n).$$

If $f(x) = +\infty$, then obviously $\lim_{n\to\infty} f_n(x) = +\infty$. Otherwise $f(x) < +\infty$ and we necessarily have $y_n \to x$. Thus

$$\lim_{n\to\infty} f_n(x) \geq \liminf_{n\to\infty} [\min(f(y_n), n) + nd(x, y_n)] \geq f(x),$$

where in the last inequality we used the fact that f is lsc. We conclude that

$$\lim_{n\to\infty} f_n(x) = f(x).$$

2. Any lsc function $f : X \to [0, +\infty]$ on a Polish space X is Borel measurable. This follows easily from the previous remark and monotone convergence.

Proof of Theorem 2.1.5. First we notice that $\mu \times \nu$ is a transference plan and therefore $\Pi(\mu, \nu) \neq \emptyset$. Since μ and ν are probability measures on a Polish space, they are tight. So for any $\epsilon > 0$ there are compact sets $K_\epsilon \subseteq X$ and $L_\epsilon \subseteq Y$ such that $\mu(X \backslash K_\epsilon) < \epsilon$ and $\nu(Y \backslash L_\epsilon) < \epsilon$. Thus we get

$$\begin{aligned}\gamma\left((X \times Y) \backslash (K_\epsilon \times L_\epsilon)\right) &\leq \gamma\left((X \backslash K_\epsilon) \times Y\right) + \gamma\left(X \times (Y \backslash L_\epsilon)\right) \\ &= \mu(X \backslash K_\epsilon) + \nu(Y \backslash L_\epsilon) \leq 2\epsilon,\end{aligned}$$

for all $\gamma \in \Pi(\mu, \nu)$. Since $K_\epsilon \times L_\epsilon$ is compact in $X \times Y$, we have that $\Pi(\mu, \nu)$ is tight. Now let $(\gamma_i)_{i \in \mathbb{N}}$ be a minimizing sequence, i.e.

$$I_c = \lim_{n \to \infty} I_c(\gamma_i).$$

Since $\Pi(\mu, \nu)$ is tight, there exists a subsequence $(\gamma_k)_{k \in \mathbb{N}}$ which converges weakly to a probability measure $\gamma^* \in \mathcal{P}(X \times Y)$. γ^* is a transference plan, because for any Borel set $A \subseteq X$ we have

$$\gamma^*(A \times Y) = \lim_{k \to \infty} \gamma_k(A \times Y) = \mu(A)$$

and analogously for any Borel $B \subseteq Y$ we have $\gamma^*(X \times B) = \nu(B)$. Hence $\gamma^* \in \Pi(\mu, \nu)$. Since $c : X \times Y \to [0, +\infty]$ is lsc, we may assume c to be the supremum of a non-decreasing sequence $(c_\ell)_{\ell \in \mathbb{N}}$ of continuous non-negative functions (see Remark 2.1.9). Thus, by weak and monotone convergence, we get

$$I_c(\gamma^*) = \int c d\gamma^* = \lim_{l \to \infty} \int c_\ell d\gamma^* = \lim_{l \to \infty} \lim_{k \to \infty} \int c_\ell d\gamma_k \leq \liminf_{k \to \infty} \int c d\gamma_k = \lim_{k \to \infty} I_c(\gamma_k),$$

hence γ^* is an optimal transference plan. □

Remark 2.1.10. We know that the MKP always has a solution (see Theorem 2.1.5). It is an open question whether the existence of optimal transference plans in the MKP can be linked to the existence of optimal transport plans for the MP.
It is easy to see that if an optimal transference plan γ^* is concentrated on the graph of a map T^*, this map is an optimal transport plan. In the next section we will study the Kantorovich duality, which will prove to be an important tool in answering this question.

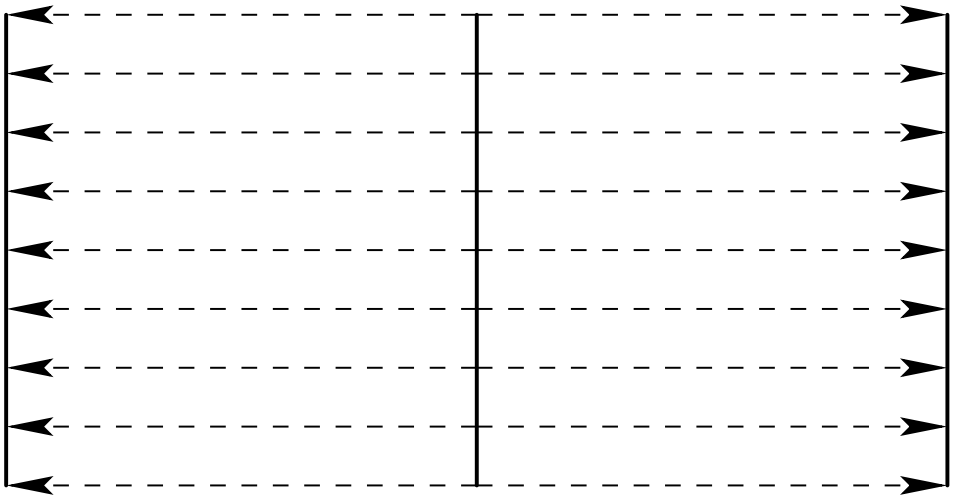


Figure 2.4: Sketch of the optimal transference plan γ^* of Example 2.1.12.

Remark 2.1.11. Let (X, d) be a Polish probability space. For $p \in [0, +\infty)$ we define a cost function $c_p(x, y) = d(x, y)^p$ and

$$W_p(\mu, \nu) := \left(\inf_{\gamma \in \Pi(\mu,\nu)} I_{c_p}(\gamma) \right)^{\frac{1}{p}}.$$

So $W_p(\mu, \nu)^p$ is the optimal transport cost between μ and ν in the MKP. One can show, that W_p is a metric on the space of probability measures on X and that it metricizes the weak convergence, i.e. $\mu_n \xrightarrow{\text{w}} \nu$ if and only if $\lim_{n\to\infty} W_p(\mu_n, \nu) = 0$.
W_p is called the **Wasserstein distance** and W_1 is also known as the **Kantorovich-Rubinstein distance**. The Wasserstein distance has very nice properties. For instance, the Kantorovich duality (see Theorem 2.2.1) implies that

$$W_p(\mu, \nu) = \left(\sup_{(\varphi,\psi) \in \Phi_{c_p}} J(\varphi, \psi) \right)^{\frac{1}{p}}$$

thus W_p may be approximated from below by some pair $(\varphi, \psi) \in \Phi_{c_p}$. We refer to Chapter 6 in [Vil09] for a thorough study.

Example 2.1.12 (Solvability MKP)**.** Let us recall the Example 2.1.2(b). Let

$$X = [0, 1] \times \{0\} \text{ and } Y = [0, 1] \times \{\pm 1\}.$$

Let μ be uniformly distributed on X and ν be uniformly distributed on Y. Further, let

the cost function be $c((x,0),(y,z)) = (x-y)^2+1$. A transference plan is given by

$$\gamma((A\times\{0\})\times[(B_1\times\{-1\})\cup(B_2\times\{1\})]) := \tfrac{1}{2}\mathcal{L}^1(A\cap B_1)+\tfrac{1}{2}\mathcal{L}^1(A\cap B_2),$$

for all measurable A, B_1, $B_2\subseteq[0,1]$. γ is indeed a transference plan because

$$\gamma((A\times\{0\})\times Y) = \mathcal{L}^1(A) = \mu(A\times\{0\})$$

and

$$\begin{aligned}\gamma(X\times((B_1\times\{-1\})\cup(B_2\times\{1\}))) &= \tfrac{1}{2}\mathcal{L}^1(B_1)+\tfrac{1}{2}\mathcal{L}^1(B_2)\\ &= \nu((B_1\times\{-1\})\cup(B_2\times\{1\}))\,.\end{aligned}$$

The cost of γ is given by

$$I_c(\gamma) = \tfrac{1}{2}\int_0^1 c((x,0),(x,-1))dx+\tfrac{1}{2}\int_0^1 c((x,0),(x,1))dx = 1.$$

In 2.1.2(b) we constructed a sequence of transport plans that converged in cost to 1. We also claimed that there was no transport plan $T\in\mathcal{T}(\mu,\nu)$ such that $I_c(T)=1$, i.e. the MP had no solution. We have now constructed a transference plan γ with $I_c(\gamma)=1$. The question that still remains is whether γ is optimal or not, i.e. $I_c(\gamma)=I_c$. At the end of the next section we will be able to answer this.

2.2 The Kantorovich Duality

The MKP has a dual representation which was introduced by Kantorovich in 1942.

Theorem 2.2.1 (Kantorovich Duality, dual MKP)**.** *Let X, Y be Polish probability spaces and μ, ν probability measures on X respectively Y. Let $c : X \times Y \to [0, +\infty]$ be a lsc cost function.*
For $\gamma \in \Pi(\mu, \nu)$ we define

$$I_c(\gamma) := \int_{X \times Y} c(x, y) d\gamma(x, y)$$

and for $(\varphi, \psi) \in \mathrm{L}^1(X, \mu) \times \mathrm{L}^1(Y, \nu)$ we define

$$J(\varphi, \psi) := \int_X \varphi d\mu + \int_Y \psi d\nu.$$

Also, define

$$\Phi_c := \{(\varphi, \psi) \in \mathrm{L}^1(X, \mu) \times \mathrm{L}^1(Y, \nu) | \varphi(x) + \psi(y) \leq c(x, y)\}.$$

Then

$$\inf_{\gamma \in \Pi(\mu, \nu)} I_c(\gamma) = \sup_{(\varphi, \psi) \in \Phi_c} J(\varphi, \psi). \tag{2.4}$$

*The problem of finding a maximizing pair $(\varphi, \psi) \in \Phi_c$ is the **dual MKP**.*

Remark 2.2.2. The Kantorovich duality actually holds under much weaker assumptions on the cost function c. In [BS11] Beiglböck and Schachermayer showed that the duality holds if c is Borel measurable, $\mu \times \nu$-a.e. finite and if there exists at least one finite transference plan.

The Kantorovich duality has a nice informal interpretation. We will use our bakery example from the beginning: X is the space of bakeries with the "density of production" μ and Y is the space of cafés with the "density of consumption" ν. Furthermore, the cost of delivering bread from a bakery $x \in X$ to a café $y \in Y$ is given by $c(x, y)$. Our task is to find an optimal transference plan $\gamma \in \Pi(\mu, \nu)$.
Suppose a company shows up with the following offer: They will pick up the bread from each bakery and deliver it to the cafés. They will charge the cost $\varphi(x)$ for each

bakery $x \in X$ and the cost $\psi(y)$ for each café $y \in Y$. Furthermore, they claim that they will never charge more than the cost we would have to pay when delivering the bread ourselves from a bakery x to a café y, i.e.

$$\varphi(x) + \psi(y) \leq c(x,y).$$

If we accept this offer we will have to pay the total cost $J(\varphi,\psi)$, which is given by

$$J(\varphi,\psi) = \int_X \varphi d\mu + \int_Y \psi d\nu.$$

Now a quick calculation shows that

$$J(\varphi,\psi) = \int_{X\times Y} \varphi(x) + \psi(y) d\gamma(x,y) \leq \int_{X\times Y} c d\gamma = I_c(\gamma).$$

So whatever transference plan $\gamma \in \Pi(\mu,\nu)$ we choose, the cost $I_c(\gamma)$ is greater or equal to $J(\varphi,\psi)$.
The Kantorovich duality tells us that, if the company also maximizes their profit, we will in fact pay as much as if we would have handled the task ourselves optimally.

To prove the Kantorovich duality we will need some preliminary results.

Assume c to be a continuous cost function. In the dual MKP one is looking for pairs (φ,ψ) of measurable functions $\varphi : X \to \mathbb{R} \cup \{-\infty\}$, $\psi : Y \to \mathbb{R} \cup \{-\infty\}$ (not equal $-\infty$ everywhere) with $\varphi(x) + \psi(y) \leq c(x,y)$, such that $J(\varphi,\psi)$ is maximal. Now, given a pair (φ,ψ), we can take a closer look at condition (2.4) and notice that for all $x \in X$ we have

$$\varphi(x) \leq c(x,y) - \psi(y). \qquad (\star)$$

Thus, if we want to increase (φ,ψ), it makes sense to try and improve φ by defining $\psi^c : X \to \mathbb{R} \cup \{-\infty\}$ by

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

Since $\psi \not\equiv -\infty$, there is a $y_0 \in Y$ such that $\psi(y_0) > -\infty$ and therefore

$$\psi^c(x) \leq c(x,y_0) - \psi(y_0) < \infty, \qquad \forall x \in X.$$

By ($\star$) $\psi^c(x) \geq \varphi(x)$, thus $\psi^c \not\equiv -\infty$ and by definition

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

Proceeding in the same way with ψ, we define $\psi^{cc} : Y \to \mathbb{R} \cup \{-\infty\}$ by

$$\psi^{cc}(y) := \inf_{x \in X} c(x,y) - \psi^c(x),$$

and again we get $\psi^{cc} \not\equiv -\infty$, $\psi^{cc} \geq \psi$ and

$$c(x,y) \geq \psi^c(x) + \psi^{cc}(y) \geq \psi^c(x) + \psi(y).$$

We still need to show, that ψ^c as well as ψ^{cc} are measurable. Since c is continuous, one can show that $-\psi^c$ is lsc and therefore Borel measurable. This can be seen in the following way: $-\psi^c$ is lsc if and only if $\mathrm{epi}(-\psi^c)$ is closed. For $y \in Y$ we define $g_y : X \to \mathbb{R}$ by $g_y(x) = \psi(y) - c(x,y)$. Then g_y is continuous, since c is continuous and $\mathrm{epi}(g_y)$ is closed. We can write

$$\mathrm{epi}(-\psi^c) = \bigcap_{y \in Y} \mathrm{epi}(g_y),$$

thus $\mathrm{epi}(-\psi^c)$ is closed and therefore $-\psi^c$ is lsc. One argues analogously to prove that $-\psi^{cc}$ is also lsc and therefore Borel measurable.
So, starting with (φ, ψ) we have constructed to a "better" pair (ψ^c, ψ^{cc}), which satisfies $\psi^c \geq \varphi$ and $\psi^{cc} \geq \psi$. We may apply the same procedure again and thus expect an even better pair (ψ^{ccc}, ψ^{cc}). But it turns out that $\psi^{ccc} = \psi^c$. This can easily be seen: First, by definition $\psi^{ccc}(x) := \inf_{y \in Y} c(x,y) - \psi^{cc}(y)$ and therefore $\psi^{ccc}(x) \geq \psi^c(x)$. On the other hand, because of $\psi(y) \leq \psi^{cc}(y)$, we get

$$\psi^{ccc}(x) \leq c(x,y) - \psi^{cc}(y) \leq c(x,y) - \psi(y)$$

and therefore

$$\psi^{ccc}(x) \leq \inf_{y \in Y} c(x,y) - \psi(y) = \psi^c(x).$$

Therefore, if we are looking for maximizing pairs in the dual problem, we can restrict

ourselves to pairs (φ, ψ) which satisfy

$$\varphi = \psi^c \text{ and } \psi = \varphi^c, \tag{2.5}$$

we will call such pairs **tight**. Given a tight pair (φ, ψ), one can always reconstruct one function from the other, so we only need to look at one of the two functions.

Definition 2.2.3 (c-transform, c-concavity and c-subdifferential)**.** Let X, Y be Polish spaces, $c : X \times Y \to [0, +\infty)$ a continuous cost function and $\varphi : X \to \mathbb{R} \cup \{-\infty\}$, $\psi : Y \to \mathbb{R} \cup \{-\infty\}$ functions, not identically $-\infty$. The functions $\varphi^c : Y \to \mathbb{R} \cup \{-\infty\}$ and $\psi^c : X \to \mathbb{R} \cup \{-\infty\}$ defined by

$$\varphi^c(y) := \inf_{x \in X} c(x, y) - \varphi(x),$$
$$\psi^c(x) := \inf_{y \in Y} c(x, y) - \psi(y),$$

are called the **c-transforms** of φ and ψ respectively.
A function φ is called **c-concave** if $\varphi^{cc} = \varphi$.
For a function φ we define

$$\partial^c \varphi := \{(x, y) \in X \times Y | \varphi(x) + \varphi^c(y) = c(x, y)\}, \tag{2.6}$$

the **c-superdifferential** of φ. Further, the c-superdifferential of φ in $x \in X$ is

$$\partial^c \varphi(x) := \{y \in Y | (x, y) \in \partial^c \varphi\} \tag{2.7}$$

or equivalently $y \in \partial_c \varphi(x)$ if and only if

$$\varphi(z) \leq \varphi(x) + c(z, y) - c(x, y) \qquad \forall z \in X.$$

Remark 2.2.4. Let c be a continuous cost function and $\psi : Y \to \mathbb{R} \cup \{-\infty\}$.

1. If there is an $x_0 \in X$ and an $\alpha \in \mathbb{R}$ such that $\psi(y) \leq c(x_0, y) - \alpha$ for all $y \in Y$, then $\psi^c(x_0) > -\infty$ and therefore $\psi^c \not\equiv -\infty$. Thus $-\psi^c$ is lsc (see the remarks before Definition 2.2.3).

2. If ψ is c-concave, then $\psi(x) + \psi^c(y) \leq c(x, y)$ and therefore $-\psi^c$ as well as $-\psi^{cc} = -\psi$ are lsc and hence Borel measurable.

Furthermore we have

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

This is completely analogous to Theorem 1.2.11 in the convex case.

3. Let ψ be c-concave and define $\tau_x : Y \to \mathbb{R}$ by

$$\tau_x(y) = c(x,y) - \psi^c(y),$$

then $\psi \leq \tau_x$ for all $x \in X$. One can think of τ_x as a tool shaped by $c(x,.)$ and $\psi \leq \tau_x$ implies that the graph of ψ can be caressed from above by such tools. In fact, a function ψ is c-concave if and only if its graph can be caressed from above by the cost function c in such a way (see figure 2.5).

4. The c-transform is a generalization of classic conjugates (Legendre-Fenchel Transform) for concave functions (see Definition 1.15). In fact, for $X = Y = \mathbb{R}^n$ and the cost function $c(x,y) = \frac{1}{2}\|x-y\|^2$, we have

$$\varphi^c(y) = \inf_{x \in \mathbb{R}^n} c(x,y) - \varphi(x) = \inf_{x \in \mathbb{R}^n} \tfrac{1}{2}\|y\|^2 - \langle x,y\rangle - (\varphi(x) - \tfrac{1}{2}\|x\|^2),$$

hence

$$\tfrac{1}{2}\|y\|^2 - \varphi^c(y) = \sup_{x \in \mathbb{R}} \langle x,y\rangle - (\tfrac{1}{2}\|x\|^2 - \varphi(x)).$$

Thus, for $h(x) := \frac{1}{2}\|x\|^2 - \varphi(x)$ we have that $h^*(y) = \frac{1}{2}\|y\|^2 - \varphi^c(y)$. So for the quadratic cost function $c(x,y) = \frac{1}{2}\|x-y\|^2$, a function φ is c-concave if and only if $\frac{1}{2}\|x\|^2 - \varphi(x)$ is convex in the usual sense.

We will need one more concept before we can prove the Kantorovich duality

Definition 2.2.5 (c-cyclical monotonicity)**.** Let X and Y be Polish spaces and $c : X \times Y \to [0, +\infty)$ a cost function. We call a set $\Gamma \subseteq X \times Y$ **c-cyclically monotone** if for all finite collections of pairs (x_1,y_1), (x_2,y_2), …, $(x_n,y_n) \in \Gamma$ we have

$$\sum_{i=1}^{n} c(x_i,y_i) \leq \sum_{i=1}^{n} c(x_i,y_{i+1}), \tag{2.8}$$

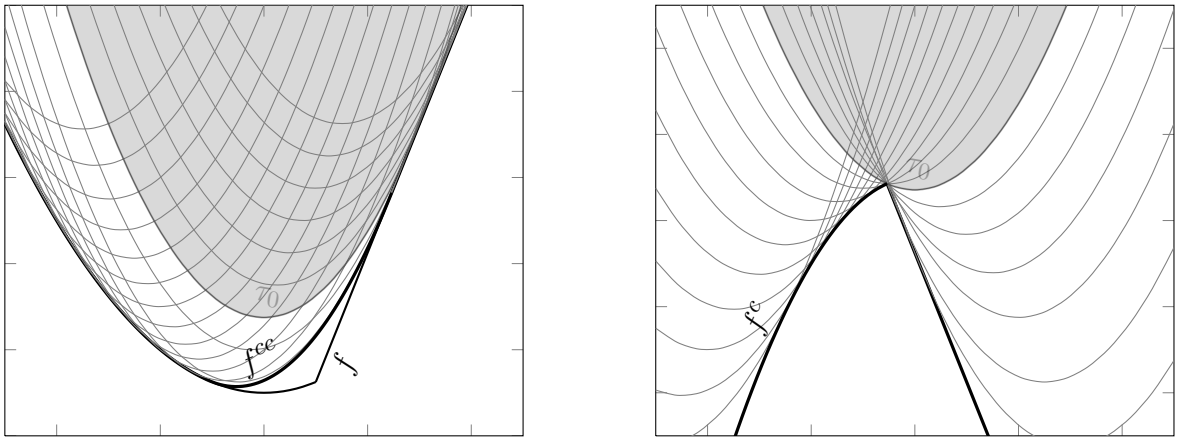


Figure 2.5: The graph of a c-concave function can be caressed from above by a tool shaped like the cost function c.

where $y_{n+1} := y_1$. Further, we will call a transference plan γ c-cyclically monotone if its **support** sptγ, defined by

$$\text{spt}\gamma := \{(x,y) \in X \times Y | \gamma(B_\epsilon(x,y)) > 0, \forall \epsilon > 0\}, \tag{2.9}$$

is c-cyclically monotone.

c-cyclical monotonicity of a transference plan is closely linked to optimality.

Theorem 2.2.6 (Finite optimal transference plans are c-cyclically monotone)**.** *Let X and Y be Polish probability spaces and μ, ν probability measures on X respectively Y. Further, let $c : X \times Y \to [0, +\infty)$ be a continuous cost function. Then a finite optimal transference plan $\gamma \in \Pi(\mu,\nu)$ is c-cyclically monotone.*

Proof. The following proof is taken from the proof of Theorem 2.3 in [GM96]. First we recall the following results from probability theory. Given a collection of measures $\mu_j \in \mathcal{P}(X)$ $(j = 1, 2, \ldots, n)$, there exists a probability space $(\Omega, \mathcal{B}(\Omega), \eta)$ such that each μ_j can be represented as the push-forward of η through a Borel map $\pi_j : \Omega \to X$. For instance, let $\eta := \mu_1 \times \mu_2 \times \ldots \times \mu_n$ be the product measure on the Borel subsets of $\Omega := X^n$, and take $\pi_j(x_1, x_2, \ldots, x_n) = x_j$ the projection onto the j-th component of Ω. Also, recall that if $U \subseteq X$ is a Borel set of mass $\mu(U) > 0$, one can define $\mu \restriction_U$, the

normalized restriction of μ to U, i.e. the probability measure defined by

$$\mu \restriction_U (V) := \frac{1}{\mu(U)}\mu(V \cap U).$$

Now, suppose that $\gamma \in \Pi(\mu,\nu)$ is optimal. We will prove indirectly that sptγ is a c-cyclically monotone set. If sptγ is not a c-cyclically monotone set, then there is an $n \in \mathbb{N}$ such that the function

$$f(x_1, x_2, \ldots, x_n; y_1, y_2, \ldots, y_n) := \sum_{i=1}^{n} c(x_i, y_{i+1}) - c(x_i, y_i)$$

is negative for some points $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n) \in \text{spt}\gamma$. Thus, since c is continuous, f is continuous and hence there are compact neighborhoods $U_j \subseteq X$ of x_j and $V_j \subseteq Y$ of y_j such that $f(u_1,\ldots,u_n;v_1,\ldots,v_n) < 0$ whenever $u_j \in U_j$ and $v_j \in V_j$. Further, since $(x_j,y_j) \in \text{spt}\gamma$, we have $\gamma(U_j \times V_j) > 0$ and therefore we can define γ_j to be the normalized restriction of γ to $U_j \times V_j$, i.e. $\gamma_j = \gamma \restriction_{U_j \times V_j}$. Also, define

$$\lambda := \min_{j=1,\ldots,n} \gamma(U_j \times V_j).$$

Then $\lambda > 0$ and

$$\gamma_j(A) = \frac{1}{\gamma(U_j \times V_j)}\gamma\left(A \cap (U_j \times V_j)\right) \leq \frac{1}{\lambda}\gamma(A),$$

which implies

$$\frac{\lambda}{n}\sum_{j=1}^{n}\gamma_j(A) \leq \gamma(A).$$

Therefore

$$\gamma - \frac{\lambda}{n}\sum_{j=1}^{n}\gamma_j$$

is a positive measure.
There is a probability space $(\Omega, \mathcal{B}(\Omega), \eta)$ and Borel maps $\omega \mapsto (\alpha_j(\omega), \beta_j(\omega))$ that take their values on $U_j \times V_j$ such that $\gamma_j = (\alpha_j \times \beta_j)\#\eta$. Thus we can define

$$\gamma' := \gamma + \frac{\lambda}{n}\sum_{j=1}^{n}(\alpha_j \times \beta_{j+1})\#\eta - (\alpha_j \times \beta_j)\#\eta.$$

Then γ' is a positive measure. Furthermore, for measurable $A \subseteq X$ we have

$$\begin{aligned}\gamma'(A\times Y) &= \gamma(A\times Y) + \frac{\lambda}{n}\sum_{j=1}^{n}(\alpha_j\times\beta_{j+1})\#\eta(A\times Y) - (\alpha_j\times\beta_j)\#\eta(A\times Y)\\ &= \mu(A) + \frac{\lambda}{n}\sum_{j=1}^{n}\eta(\alpha_j^{-1}(A)) - \eta(\alpha_j^{-1}(A)) = \mu(A)\end{aligned}$$

and analogously we get $\gamma'(X\times B) = \nu(B)$, hence $\gamma' \in \Pi(\mu,\nu)$. Using the fact that f is negative on $(\prod_{j=1}^{n} U_j)\times(\prod_{j=1}^{n} V_j)$ and that α_j takes values only in U_j and β_j only in V_j, we can conclude that

$$I_c(\gamma') - I_c(\gamma) = \frac{\lambda}{n}\int_{\Omega}\sum_{j=1}^{n} c(\alpha_j,\beta_{j+1}) - c(\alpha_j,\beta_j)d\eta < 0.$$

Therefore $I_c(\gamma') < I_c(\gamma)$, which is a contradiction to the optimality of γ. Thus sptγ is c-cyclically monotone. □

Remark 2.2.7. Actually a lot more is true. One can show that for a Borel measurable cost function $c : X\times Y \to [0,+\infty]$ every finite optimal transference plan $\gamma \in \Pi(\mu,\nu)$ is concentrated on a c-cyclically monotone set. Also, if a finite transference plan is concentrated on a c-cyclically monotone set it is optimal if $\{c = +\infty\} = F \cup N$, where $F \subseteq X\times Y$ is closed and N is a $\mu\times\nu$-null set. These results can be found in [BGMS09] Theorem 1.

Theorem 2.2.8 (Characterization of c-cyclically monotone sets)**.** *Let X and Y be Polish spaces and $c : X\times Y \to [0,+\infty)$ be a continuous cost function. A set $\Gamma \subseteq X\times Y$ is c-cyclically monotone if and only if there exists a c-concave function φ such that $\Gamma \subseteq \partial^c\varphi$.*

Proof. Let φ be c-concave. We will first show that $\partial^c\varphi$ is c-cyclically monotone. Let $n \in \mathbb{N}$ and $(x_1,y_1), \ldots, (x_n,y_n) \in \partial^c\varphi$. Then, by definition of $\partial^c\varphi$, we have for all $z \in X$

$$\varphi(z) \le \varphi(x_j) + c(z,y_j) - c(x_j,y_j),$$

for $j = 1,\ldots,n$. Choosing $z = x_{j-1}$ we can write

$$\varphi(x_{j-1}) - \varphi(x_j) \le c(x_{j-1},y_j) - c(x_j,y_j),$$

hence we get

$$\sum_{j=1}^{n} c(x_j, y_{j+1}) - c(x_j, y_j) \geq \sum_{j=1}^{n} \varphi(x_{j-1}) - \varphi(x_j) = 0.$$

Therefore $\partial^c\varphi$ is c-cyclically monotone and thus also all subsets of $\partial^c\varphi$ are c-cyclically monotone.
To prove necessity assume $\Gamma \subseteq X \times Y$ to be c-cyclically monotone. We need to construct a c-concave function φ such that $\Gamma \subseteq \partial^c\varphi$. Fix $(x_0, y_0) \in \Gamma$ and define $\varphi_n : X \times \Gamma^n \to \mathbb{R}$ by

$$\varphi_n(x; x_1, y_1, \ldots, x_n, y_n) := [c(x, y_n) - c(x_n, y_n)] + \sum_{i=0}^{n-1} [c(x_{i+1}, y_i) - c(x_i, y_i)].$$

Next, define $\varphi : X \to \mathbb{R} \cup \{-\infty\}$ by

$$\varphi(x) := \inf\{\varphi_n(x; x_1, y_1, \ldots x_n, y_n) | n \in \mathbb{N}, (x_1, y_1), \ldots, (x_n, y_n) \in \Gamma\}.$$

Using the definition of φ for $n = 1$ and $(x_1, y_1) = (x_0, y_0)$ we see that $\varphi(x_0) \leq 0$. On the other hand for $x = x_0$ we have, using c-cyclical monotonicity of Γ,

$$\varphi_n(x_0; x_1, y_1, \ldots, x_n, y_n) = \sum_{i=0}^{n} c(x_{i+1}, y_i) - c(x_i, y_i) \geq 0,$$

thus $\varphi(x_0) = 0$.
Relabeling $y_n = y$ we get

$$\varphi_n(x; x_1, y_1, \ldots x_n, y) = c(x, y) - c(x_n, y) + \sum_{i=0}^{n-1} [c(x_{i+1}, y_i) - c(x_i, y_i)],$$

hence defining $\psi_n : Y \times \Gamma^{n-1} \times \Gamma_1 \to \mathbb{R}$ by

$$\psi_n(y; x_1, y_1, \ldots x_{n-1}, y_{n-1}, x_n) := c(x_n, y) - \sum_{i=0}^{n-1} [c(x_{i+1}, y_i) - c(x_i, y_i)],$$

where $\Gamma_1 = \{x | \exists y \in Y : (x, y) \in \Gamma\}$, we obtain

$$\varphi_n(x; x_1, y_1, \ldots x_n, y) = c(x, y) - \psi_n(y; x_1, y_1, \ldots x_{n-1}, y_{n-1}, x_n).$$

Therefore, for $\psi : Y \to \mathbb{R} \cup \{\pm\infty\}$ defined by

$$\psi(y) := \sup\{\psi_n(y; x_1, y_1, \dots x_{n-1}, y_{n-1}, x_n) | n \in \mathbb{N}, (x_1, y_1), \dots, (x_n, y) \in \Gamma\},$$

for $y \in \Gamma_2 = \{y | \exists x \in X : (x, y) \in \Gamma\}$ and $\psi(y) := -\infty$ otherwise, we get

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

Also, since $\varphi(x_0) = 0$, we have that $\psi(y) < +\infty$ for all $y \in Y$. Thus

$$\varphi^{cc}(x) = \psi^{ccc}(x) = \psi^c(x) = \varphi(x)$$

and therefore φ is c-concave.
Now let $(\overline{x}, \overline{y}) \in \Gamma$. We can write

$$\varphi_n(z; x_1, y_1, \dots, x_{n-1}, y_{n-1}, \overline{x}, \overline{y}) = \varphi_{n-1}(\overline{x}; x_1, y_1, \dots, x_{n-1}, y_{n-1}) + c(z, \overline{y}) - c(\overline{x}, \overline{y}),$$

thus

$$\varphi(z) \leq \varphi(\overline{x}) + c(z, \overline{y}) - c(\overline{x}, \overline{y})$$

for all $z \in X$. Therefore $\overline{y} \in \partial^c \varphi(\overline{x})$, which implies $\Gamma \subseteq \partial^c \varphi$. □

We can now proof the Kantorovich duality.

Proof of 2.2.1. First we notice that for any pair $(\varphi, \psi) \in \Phi_c$ and $\gamma \in \Pi(\mu, \nu)$ we have

$$J(\varphi, \psi) = \int_X \varphi d\mu + \int_Y \psi d\nu = \int_{X \times Y} \varphi(x) + \psi(y) d\gamma(x, y) \leq \int_{X \times Y} c d\gamma = I_c(\gamma),$$

hence $J(\varphi, \psi) \leq I_c$. Therefore we only need to show the inequality

$$\sup_{(\varphi, \psi) \in \Phi_c} J(\varphi, \psi) \geq I_c.$$

We will assume c to be bounded and continuous. Then, because of Theorem 2.1.5, there

is always a solution $\gamma^* \in \Pi(\mu,\nu)$ to the MKP and because

$$I_c = I_c(\gamma^*) = \int_{X\times Y} c d\gamma^* \leq \sup\{c(x,y)|(x,y)\in X\times Y\} < +\infty,$$

γ^* is finite. Thus, because of Theorem 2.2.6, γ^* has c-cyclically monotone support $\Gamma := \text{spt}\gamma^*$. So, using Theorem 2.2.8, we get a c-concave, and therefore measurable, function φ such that $\Gamma \subseteq \partial^c\varphi$. For $(x,y)\in\partial^c\varphi$ we have by definition $\varphi(x)+\varphi^c(y) = c(x,y)$ and therefore we can conclude that

$$J(\varphi,\varphi^c) = \int_{X\times Y} \varphi(x)+\varphi^c(y) d\gamma(x,y) = \int_{X\times Y} c d\gamma = I_c(\gamma) = I_c.$$

Thus we have proved the Kantorovich duality for a bounded and continuous cost function. Now if c is lsc, then c can be written as the supremum of a non-decreasing sequence $(c_k)_{k\in\mathbb{N}}$ of bounded and continuous functions (see Remark 2.1.9). We will show that $I_{c_k} \to I_c$. Since $c_k \leq c$, we have

$$I_{c_k}(\gamma) \leq I_c(\gamma) \text{ for all } \gamma\in\Pi(\mu,\nu),$$

hence $I_{c_k} \leq I_c$. Now, for each $k\in\mathbb{N}$ let $\gamma_k \in \Pi(\mu,\nu)$ be optimal for c_k, i.e. $I_{c_k}(\gamma_k) = I_{c_k}$. Since $\Pi(\mu,\nu)$ is tight, there is a subsequence $(\gamma_\ell)_{\ell\in\mathbb{N}}$ that converges weakly to a $\gamma\in\Pi(\mu,\nu)$. We conclude

$$I_c \leq I_c(\gamma) = \lim_{k\to\infty}\int_{X\times Y} c_k d\gamma = \lim_{k\to\infty}\lim_{\ell\to\infty}\int_{X\times Y} c_k d\gamma_\ell \leq \lim_{k\to\infty}\lim_{\ell\to\infty}\int_{X\times Y} c_\ell d\gamma_\ell = \lim_{\ell\to\infty} I_{c_\ell}(\gamma_\ell).$$

Hence $I_{c_k}\to I_c$ as $k\to\infty$. Next, for every k we may assume $(\varphi_k,\psi_k)\in\Phi_{c_k}$ to satisfy

$$J(\varphi_k,\psi_k) \geq I_{c_k} - \tfrac{1}{k}.$$

Since $c_k \leq c$ we have $(\varphi_k,\psi_k)\in\Phi_c$ and thus

$$\lim_{k\to\infty} J(\varphi_k,\psi_k) \geq \lim_{k\to\infty} I_{c_k} - \tfrac{1}{k} = I_c.$$

Therefore we have proved the Kantorovich duality for lsc cost functions. □

Corollary 2.2.9 (Maximizers for the dual MKP)**.** *Let X, Y be Polish probability spaces and μ, ν probability measures on X respectively Y. Let $c: X\times Y\to[0,+\infty]$ be a*

continuous cost function. If either

(i) there exists a finite optimal transference plan $\gamma \in \Pi(\mu,\nu)$, or

(ii) c is bounded from above

then there exists $\varphi : X \to \mathbb{R} \cup \{-\infty\}$ c-concave such that $(\varphi, \varphi^c) \in \Phi_c$ maximizes the dual MKP.

Proof. In case (i), by Theorem 2.2.6, sptγ is c-cyclically monotone. Thus, by Theorem 2.2.8 there is $\varphi : X \to \mathbb{R} \cup \{-\infty\}$ c-concave such that spt$\gamma \subseteq \partial^c\varphi$. Thus $I_c(\gamma) = J(\varphi, \varphi^c)$ and therefore (φ, φ^c) is a maximizing pair of the dual MKP.
In case (ii), for $\gamma \in \Pi(\mu,\nu)$ we have

$$I_c(\gamma) \leq \sup_{(x,y)\in X\times Y} c(x,y) < +\infty,$$

thus an optimal transference plan (which exists by Theorem 2.1.5) is finite and we can apply (i) to complete this proof. □

Corollary 2.2.10 (c-cyclically monotone transference plans are optimal)**.** *Let X, Y be Polish spaces and μ, ν probability measures on X respectively Y. Let $c : X \times Y \to [0, +\infty]$ be a continuous cost function. If a transference plan $\gamma \in \Pi(\mu,\nu)$ has c-cyclically monotone support, then it is optimal.*

Proof. Since γ has c-cyclically monotone support, by Theorem 2.2.8, there is a c-concave function φ (which is measurable since c is continuous and therefore φ is upper semi-continuous) such that spt$\gamma \subseteq \partial^c\varphi$, thus $\varphi(x) + \varphi^c(y) = c(x,y)$ γ-almost everywhere. Hence we get

$$I_c \geq J(\varphi, \varphi^c) = I_c(\gamma),$$

Therefore γ is optimal. □

Example 2.2.11 (Continuation of Example 2.1.12)**.** We recall the previous example. Let

$$X = [0,1] \times \{0\} \text{ and } Y = [0,1] \times \{\pm 1\}.$$

Let μ be uniformly distributed on X and ν be uniformly distributed on Y. Furthermore, let the cost function be $c((x,0),(y,z)) = (x-y)^2+1$. We have previously constructed a transference plan $\gamma \in \Pi(\mu,\nu)$ such that $I_c(\gamma) = 1$. We will now use the Kantorovich duality to prove that γ is optimal. Let $\varphi : X \to \mathbb{R}$ and $\psi : Y \to \mathbb{R}$ defined by $\varphi(x) = \frac{1}{2}$ and $\psi(y) = \frac{1}{2}$. Then $\varphi(x)+\psi(y) = 1 \leq c(x,y)$ and

$$J(\varphi,\psi) = \int_X \varphi d\mu + \int_Y \psi d\nu = \tfrac{1}{2}\mu(X) + \tfrac{1}{2}\nu(Y) = 1 = I_c(\gamma).$$

Thus γ is optimal and (φ,ψ) is a maximizing pair in the dual Problem. We have also just proved the claim we made in 2.1.2 that the optimal cost of the Monge Problem is equal to $I_c = 1$.

2.3 Optimal Transport on the Real Line

In this section we consider the MKP on the real line. The results we obtain will help us, when we treat the MKP on the circle. On the real line the Optimal Transport Problem can be solved explicitly for cost functions of the form $c(x,y) = \lambda(|x-y|)$, where λ is strictly convex, non-negative and increasing.

Definition 2.3.1 (Cumulative distribution function and generalized inverse)**.** Let μ be a probability measure on $\mathbb{R}$. $F_\mu : \mathbb{R} \to [0,1]$, defined by

$$F_\mu(x) := \mu((-\infty, x]), \tag{2.10}$$

is called the **cumulative distribution function** of μ.
The function $F_\mu^{-1} : [0,1] \to \mathbb{R} \cup \{\pm\infty\}$ defined by

$$F_\mu^{-1}(y) := \inf\{x \in \mathbb{R} | y < F_\mu(x)\} \tag{2.11}$$

is called the **generalized inverse** of F_μ.

Remark 2.3.2.

1. F_μ as well as F_μ^{-1} are right-continuous and non-decreasing.
2. One can reconstruct μ from F_μ in the following way: For $a \leq b$ we have $\mu((a,b]) = F_\mu(b) - F_\mu(a)$ and the sets $(a,b]$ generate all Borel sets on $\mathbb{R}$.

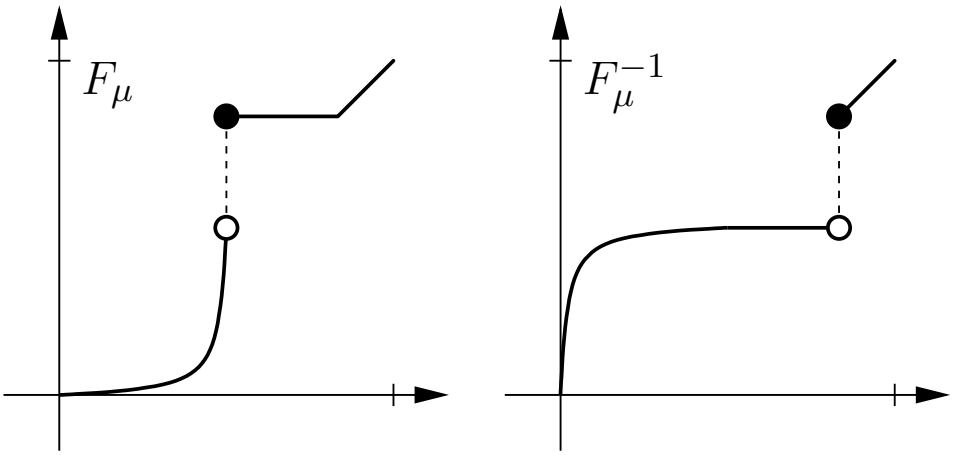


Figure 2.6: The cumulative distribution function F_μ and its generalized inverse F_μ^{-1} of a measure μ.

3. If μ is absolutely continuous with respect to the Lebesgue measure $\mathcal{L}^1$ on $\mathbb{R}$, then μ does not assign mass to points and F_μ is continuous. Further, $F_\mu(F_\mu^{-1}(t)) = t$.
4. $F_\mu^{-1}(a) = -\infty$ if and only if $a = 0$ respectively $F_\mu^{-1}(b) = +\infty$ only if $b = 1$, thus F_μ^{-1} is real valued on the open interval $(0, 1)$.
5. Sometimes the generalized inverse is defined by $F_\ell^{-1}(y) := \inf\{x \in \mathbb{R} | y \leq F_\mu(x)\}$; then F_ℓ^{-1} is left-continuous instead of right-continuous.

Definition 2.3.3 (Monotone sets)**.** We call a set $\Gamma \subseteq \mathbb{R}^2$ **monotone** if and only if for each (x_1, y_1), $(x_2, y_2) \in \Gamma$

$$(x_1 - x_2)(y_1 - y_2) \geq 0,$$

or equivalently, if $x_1 < x_2$ implies $y_1 \leq y_2$ and $y_1 < y_2$ implies $x_1 \leq x_2$.

Remark 2.3.4. Monotone sets can be considered as the complete graphs of monotone functions. For instance if $f : \mathbb{R} \to \mathbb{R}$ is increasing, then its graph is a monotone set if we add the segments $[f(x^-), f(x^+)]$ for all $x \in \mathbb{R}$ where f is discontinuous (see figure 2.7).

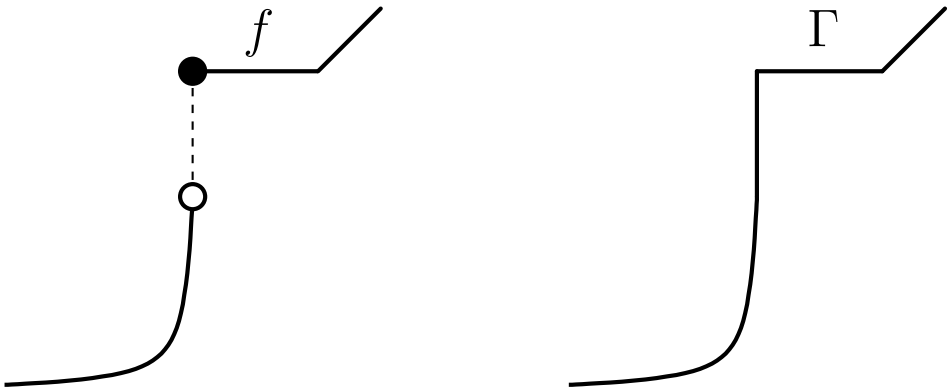


Figure 2.7: A monotone set Γ is the complete graph of a non-decreasing function f.

Lemma 2.3.5 (Monotonicity vs. c-cyclical monotonicity)**.** *Let $c(x,y) = \lambda(|x-y|)$ be a cost function.*

(i) If λ is convex, increasing and non-negative, then every monotone set $\Gamma \subseteq \mathbb{R}^2$ is c-cyclically monotone.

(ii) If λ is also strictly convex, then every c-cyclically monotone set $\Gamma \subseteq \mathbb{R}^2$ is monotone as well.

Proof.

1. Assume λ to be strictly convex. First we show that

$$c(x_1,y_1)+c(x_2,y_2) < c(x_1,y_2)+c(x_2,y_1) \qquad (\star)$$

if and only if

$$(x_1-x_2)(y_1-y_2) > 0.$$

If we assume $x_1 < x_2$ we will need to show that $(\star)$ is true exactly for $y_1 < y_2$. To prove this, we will look at all the possible alignments of the points x_1, x_2, y_1 and y_2. Let

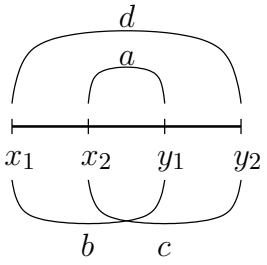
$$a := |x_2-y_1|, \qquad b := |x_1-y_1|, \qquad c := |x_2-y_2|, \qquad d := |x_1-y_2|.$$

So, to prove ($\star$), we need to show

$$\lambda(b)+\lambda(c)<\lambda(a)+\lambda(d).$$

Case (i): If $x_1 < x_2 \le y_1 < y_2$, we get

$$a<b<d \text{ and } a<c<d.$$



We see that

$$b=\frac{d-b}{d-a}a+\frac{b-a}{d-a}d,$$
$$c=\frac{d-c}{d-a}a+\frac{c-a}{d-a}d,$$

with

$$\frac{d-b}{d-a}+\frac{b-a}{d-a}=1 \text{ and } 0<\frac{d-b}{d-a}<1$$

as well as

$$\frac{d-c}{d-a}+\frac{c-a}{d-a}=1 \text{ and } 0<\frac{d-c}{d-a}<1.$$

Now, since λ is strictly convex, we get

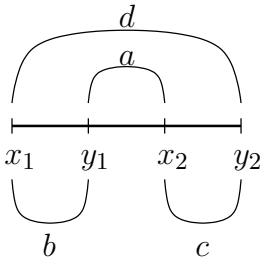
$$\lambda(b)<\frac{d-b}{d-a}\lambda(a)+\frac{b-a}{d-a}\lambda(d),$$
$$\lambda(c)<\frac{d-c}{d-a}\lambda(a)+\frac{c-a}{d-a}\lambda(d).$$

By taking the sum of these inequalities, and since $a+d=b+c$, we get

$$\lambda(b)+\lambda(c)<\lambda(a)+\lambda(d).$$

Case (ii): If $x_1 \le y_1 < x_2 \le y_2$, we get

$$a+b+c=d.$$



We see that $d>b$ and $d>c$. Since λ is strictly increasing we get $\lambda(d)>\lambda(b)$ and $\lambda(d)>\lambda(c)$. So, if either $a \ge b$ or $a \ge c$, we also get $\lambda(a) \ge \lambda(b)$ or $\lambda(a) \ge \lambda(c)$.

Therefore

$$\lambda(a)+\lambda(d)>\lambda(b)+\lambda(c).$$

Otherwise, we have $a<b,c<d$ and can use the same procedure as in case (i) to get

$$\lambda(b)<\frac{d-b}{d-a}\lambda(a)+\frac{b-a}{d-a}\lambda(d),$$
$$\lambda(c)<\frac{d-c}{d-a}\lambda(a)+\frac{c-a}{d-a}\lambda(d).$$

We obtain

$$\lambda(b)+\lambda(c)<\frac{2d-b-c}{d-a}\lambda(a)+\frac{b+c-2a}{d-a}\lambda(d).$$

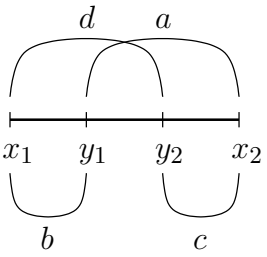
Using the equation $a+b+c=d$ we can rewrite this to

$$\begin{aligned}\lambda(b)+\lambda(c)&<\left(1+\frac{2a}{d-a}\right)\lambda(a)+\left(1-\frac{2a}{d-a}\right)\lambda(d)\\&=\lambda(a)+\lambda(d)+\frac{2a}{d-a}(\lambda(a)-\lambda(d))<\lambda(a)+\lambda(d),\end{aligned}$$

where in the last inequality we used the fact that $\lambda(a)<\lambda(d)$.

Case (iii): If $x_1\le y_1<y_2\le x_2$, we get

$$b<d \text{ and } c<a.$$



Thus $\lambda(b)+\lambda(c)<\lambda(a)+\lambda(d)$, because λ is a strictly increasing function.

Case (iv): All other cases can be reduced to one of the cases (i) - (iii) by setting $(x_1',y_1'):=(y_1,x_1)$ and $(x_2',y_2'):=(y_2,x_2)$.

2. If Γ is c-cyclically monotone, then for (x_1,y_1), $(x_2,y_2)\in\Gamma$ we have

$$c(x_1,y_1)+c(x_2,y_2)\le c(x_1,y_2)+c(x_2,y_1).$$

Using step 1. we can deduce that this either implies $x_1 = x_2$ or $y_1 = y_2$ and therefore

$$(x_1 - x_2)(y_1 - y_2) = 0,$$

or

$$(x_1 - x_2)(y_1 - y_2) > 0.$$

So c-cyclically monotone sets are monotone.

3. Now assume λ to be convex, increasing and non-negative and Γ to be monotone. We can argue as in 1. to see that

$$(x_1 - x_2)(y_1 - y_2) > 0$$

implies

$$c(x_1, y_1) + c(x_2, y_2) \leq c(x_1, y_2) + c(x_2, y_1).$$

We need to show that

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

for arbitrary (x_1, y_1), (x_2, y_2), ..., $(x_n, y_n) \in \Gamma$.
We will prove this by induction on n. For $n = 1$ we have nothing to show. Let $n > 1$ and assume that ($\star\star$) holds for $n - 1$. Let (x_1, y_1), ..., $(x_n, y_n) \in \Gamma$ be arbitrary and choose $k \in \{1, \ldots, n\}$ to be the index of the largest y_i, i.e.

$$y_k = \max_{i=1,\ldots,n} \{y_i\}.$$

We will show

$$c(x_{k-1}, y_k) + c(x_k, y_{k+1}) \geq c(x_{k-1}, y_{k+1}) + c(x_k, y_k). \tag{$\star\star\star$}$$

By definition of k we have $y_k \geq y_{k-1}$ and by using the monotonicity of Γ we obtain $x_k \geq x_{k-1}$. We also have $y_k \geq y_{k+1}$ thus

$$(x_k - x_{k-1})(y_k - y_{k+1}) \geq 0$$

which implies ($\star\star\star$).
Now put

$$(\overline{x}_i, \overline{y}_i) := \begin{cases} (x_i, y_i) & \text{if } i < k, \\ (x_{i+1}, y_{i+1}) & \text{if } i \geq k. \end{cases}$$

Then

$$\sum_{i=1}^{n} c(x_i, y_{i+1}) \geq c(x_k, y_k) + \sum_{i=1}^{n-1} c(\overline{x}_i, \overline{y}_{i+1}).$$

Using our induction hypothesis for $n-1$ yields

$$c(x_k, y_k) + \sum_{i=1}^{n-1} c(\overline{x}_i, \overline{y}_{i+1}) \geq c(x_k, y_k) + \sum_{i=1}^{n-1} c(\overline{x}_i, \overline{y}_i) = \sum_{i=1}^{n} c(x_i, y_i).$$

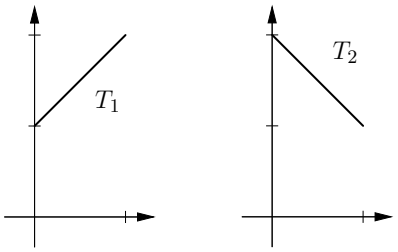
Hence ($\star\star$) holds for n, which completes our induction. Thus Γ is c-cyclically monotone. □

Example 2.3.6. Let $c(x,y) = \lambda(|x-y|)$ be a convex cost function (i.e. λ is convex, non-negative and increasing), μ the Lebesgue measure on $[0,1]$ and ν the Lebesgue measure on $[1,2]$. Consider the transport plans

$$T_1 : \begin{cases} [0,1] & \to & [1,2], \\ x & \mapsto & x+1, \end{cases}$$

and

$$T_2 : \begin{cases} [0,1] & \to & [1,2], \\ x & \mapsto & 2-x. \end{cases}$$



The support of γ_{T_1} is the graph of the function T_1. Therefore the support of γ_{T_1} is a monotone set and since λ is convex, it is also c-cyclically monotone and hence γ_{T_1} is optimal.
Let $c(x,y) = |x-y|$, i.e. c is convex but not strictly convex. Then the cost of T_1 resp. T_2

is given by

$$I_c(\gamma_{T_1}) = \int_0^1 c(t, T_1(t))dt = 1,$$
$$I_c(\gamma_{T_2}) = \int_0^1 c(t, T_2(t))dt = \int_0^1 |2t-2|dt = 1.$$

Thus T_1 as well as T_2 are optimal transport plans. Since γ_{T_2} is optimal, it has a c-cyclically monotone support. Thus the graph of the function T_2 is a c-cyclically monotone set, but it obviously is not monotone. This shows that the distinction between convex and strictly convex cost functions in Lemma 2.3.5 is necessary.
We could also use the Kantorovich Duality to show that the optimal cost is equal to 1: Choose $\varphi(x) = -x$ and $\psi(y) = y$. Then

$$\varphi(x) + \psi(y) = y - x = |x-y| = c(x,y), \qquad \forall x \in [0,1], y \in [1,2]$$

and

$$J(\varphi, \psi) = \int_0^1 \varphi(x)dx + \int_1^2 \psi(y)dy = 1.$$

This also shows that T_1 and T_2 are optimal.
If we choose $c(x,y) = |x-y|^2$, then c is a strictly convex cost function and γ_{T_1} is still optimal. Since γ_{T_2} has non-monotone support, it cannot be optimal. If it were optimal, its support would necessarily be c-cyclically monotone and therefore, since c is strictly convex, also monotone – a contradiction. And indeed the costs are

$$I_c(\gamma_{T_1}) = \int_0^1 |x-(x+1)|^2 dx = 1,$$
$$I_c(\gamma_{T_2}) = \int_0^1 |2x-2|^2 dx = 1 + \tfrac{1}{3}.$$

Thus T_2 is not optimal.

Theorem 2.3.7 (MKP on the real line for strictly convex cost function)**.**
Let μ, ν be two probability measures on $\mathbb{R}$ and $c(x,y) = \lambda(|x-y|)$ a convex cost function (i.e. λ is convex, non-negative and increasing). We define a probability measure γ^* on

$\mathbb{R}^2$ *by*

$$\gamma^* := \left(F_\mu^{-1} \times F_\nu^{-1}\right) \# \mathcal{L}^1, \tag{2.12}$$

where $\mathcal{L}^1$ *is the Lebesgue measure on* $\mathbb{R}$*. Then* $\gamma^* \in \Pi(\mu,\nu)$*, i.e.* γ^* *is a transference plan and if* $I_c(\gamma^*) < +\infty$ *then it is optimal. In this case, the optimal transport cost is given by*

$$I_c = I_c(\gamma^*) = \int\limits_{\mathbb{R}\times\mathbb{R}} c(x,y) d\gamma^*(x,y) = \int\limits_0^1 \lambda(|F_\mu^{-1}(t) - F_\nu^{-1}(t)|) dt.$$

Furthermore, if μ *is absolutely continuous with respect to* $\mathcal{L}^1$*, then*

$$T := F_\nu^{-1} \circ F_\mu \tag{2.13}$$

is an optimal transport plan, i.e. $I_c(\gamma_T) = I_c(\gamma^*)$*.*

Proof. The following proof was given by Villani in [Vil03](Theorem 2.18).

1. First we check that γ^* is a transference plan from μ to ν. For the set $R(x,y) = (-\infty, x] \times (-\infty, y]$ we have

$$\begin{aligned} \gamma^*(R(x,y)) &= \mathcal{L}^1\{t \in [0,1] | F_\mu^{-1}(t) \leq x \text{ and } F_\nu^{-1}(t) \leq y\} \\ &= \min(F_\mu(x), F_\nu(y)), \end{aligned}$$

   since $\{t \in [0,1] | F_\mu^{-1}(t) \leq x\}$ is either $[0, F_\mu(x)]$ or $[0, F_\mu(x))$ and analogously for ν. Therefore

$$\gamma^*((-\infty, x] \times \mathbb{R}) = F_\mu(x) \text{ and } \gamma^*(\mathbb{R} \times (-\infty, y]) = F_\nu(y),$$

   which implies $\gamma^*(A \times \mathbb{R}) = \mu(A)$ and $\gamma^*(\mathbb{R} \times B) = \nu(B)$ for measurable A and B, so γ^* is a transference plan.

2. We will show, for $(x,y) \in \mathrm{spt}(\gamma^*)$, that

$$F_\mu(x^-) \leq F_\nu(y) \text{ and } F_\nu(y^-) \leq F_\mu(x).$$

   Assume that $F_\mu(x^-) > F_\nu(y)$. Now, since F_ν is right-continuous and F_μ as well as F_ν are non-decreasing, for x' in a small neighborhood of x and y' in a small

neighborhood of y, we get that $F_\mu(x') > F_\nu(y')$ and therefore $\gamma^*(R(x',y')) = F_\nu(y')$. So for a sufficiently small rectangle $(x',x''] \times (y',y'']$ with $x' < x < x''$ and $y' < y < y''$ we have

$$\begin{aligned}\gamma^*((x',x''] \times (y',y'']) &= \gamma^*(R(x'',y'')) - \gamma^*(R(x',y'')) - \gamma^*(R(x'',y')) + \gamma^*(R(x',y')) \\ &= F_\nu(y'') - F_\nu(y'') - F_\nu(y') + F_\nu(y') = 0.\end{aligned}$$

Thus $(x,y) \notin \operatorname{spt}(\gamma^*)$.

3. We will now show that γ^* has c-cyclical monotone support and, since $I_c(\gamma^*)$ is finite, is therefore optimal (see Corollary 2.2.10). Let (x_1,y_1), $(x_2,y_2) \in \operatorname{spt}(\gamma^*)$ and assume that $x_1 < x_2$. Since the cost function is convex it is sufficient to show that $y_1 \leq y_2$ (see Lemma 2.3.5). We can use the previous step to deduce

$$F_\nu(y_1^-) \leq F_\mu(x_1) \leq F_\mu(x_2^-) \leq F_\nu(y_2).$$

If $F_\nu(y_1^-) < F_\nu(y_2)$, then $y_1 \leq y_2$ and we are done.
Otherwise, necessarily $F_\nu(y_1) = F_\mu(x_1^-) = F_\mu(x_2) = F_\nu(y_2^-)$. We will assume $y_1 > y_2$ and show, that (x_1,y_1) can not belong to $\operatorname{spt}(\gamma^*)$. Since $F_\mu(x_1) = F_\mu(x_2^-)$ we see that F_μ is constant on $[x_1,x_2)$ and because of $F_\nu(y_1^-) = F_\nu(y_2)$ we see that F_ν is constant on $[y_2,y_1)$. So if we consider the rectangle $(x_1-\epsilon,x_1+\epsilon] \times (y_1-\epsilon,y_1+\epsilon]$, then for ϵ small enough we get

$$F_\mu(x_1-\epsilon) \leq F_\mu(x_1+\epsilon) = F_\mu(x_1) = F_\nu(y_1^-) = F_\nu(y_1-\epsilon) \leq F_\nu(y_1+\epsilon).$$

Thus we have

$$\begin{aligned}&\gamma^*((x_1-\epsilon,x_1+\epsilon] \times (y_1-\epsilon,y_1+\epsilon]) \\ &\quad= \gamma^*(R(x_1+\epsilon,y_1+\epsilon)) - \gamma^*(R(x_1-\epsilon,y_1+\epsilon)) \\ &\quad\quad- \gamma^*(R(x_1+\epsilon,y_1-\epsilon)) + \gamma^*(R(x_1-\epsilon,y_1-\epsilon)) \\ &\quad= F_\mu(x_1+\epsilon) - F_\mu(x_1-\epsilon) - F_\mu(x_1+\epsilon) + F_\mu(x_1-\epsilon) = 0,\end{aligned}$$

and therefore $(x_1,y_1) \notin \operatorname{spt}(\gamma^*)$, which is a contradiction.

4. If μ is absolutely continuous, then F_μ is continuous and $F_\mu(F_\mu^{-1}(t)) = t$. Therefore, letting $T = F_\nu^{-1} \circ F_\mu$, we will show that $\rho := (\mathrm{id} \times T)\#\mu$ coincides with γ^*. Let

$u(x,y)$ be a measurable function on $\mathbb{R}^2$, then

$$\begin{aligned}\int u(x,y)d\rho(x,y) &= \int u(x,T(x))d\mu(x)\\ &= \int u(x,F_\nu^{-1}\circ F_\mu(x))d\mu(x)\\ &= \int u(F_\mu^{-1}(t),F_\nu^{-1}\circ F_\mu\circ F_\mu^{-1}(t))dt\\ &= \int u(F_\mu^{-1}(t),F_\nu^{-1}(t))dt = \int u(x,y)d\gamma^*(x,y),\end{aligned}$$

where we used the substitution $F_\mu^{-1}(t) = x$. □

Chapter 3

Optimal Transport on the *n*-Sphere

In this chapter we will study the MKP on the n-sphere $\mathbb{S}^n$. We will restrict ourselves to cost functions c of the form $c(x,y) = \lambda(d(x,y))$, where $d : \mathbb{S}^n \times \mathbb{S}^n \to [0,\pi]$ is the **chord length metric** on $\mathbb{S}^n$, i.e.

$$d(x,y) := \arccos(\langle x,y\rangle), \qquad \forall x,y \in \mathbb{S}^n \tag{3.1}$$

and $\lambda : [0,\pi] \to [0,+\infty)$ is a strictly convex, non-negative and increasing function. We will call such cost functions **strictly convex**. To prove our main Theorem we will also require λ to satisfy some more conditions (see Proposition 3.4.2).

Let c be a strictly convex cost function and μ, ν probability measures on $\mathbb{S}^n$. Then c is obviously bounded and continuous. By Theorem 2.1.5 there is an optimal transference plan $\gamma^* \in \Pi(\mu,\nu)$ and by Corollary 2.2.9 there is also a c-concave function $\varphi : \mathbb{S}^n \to \mathbb{R}$ that maximizes the dual MKP, thus

$$I_c(\gamma^*) = J(\varphi, \varphi^c).$$

In the sequel we will show that if μ is absolutely continuous with respect to σ^n, where σ^n denotes the **volume measure** on $\mathbb{S}^n$, then there is a σ^n-a.e. uniquely determined optimal transport plan $T : \mathbb{S}^n \to \mathbb{S}^n$ that solves the MKP.

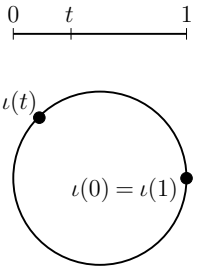
But first we will consider the 1-dimensional case of the circle.

3.1 Optimal Transport on the Circle

In this section we will study the MKP on the circle for strictly convex cost functions. We will use the results of Theorem 2.3.7. To do so, we will basically 'cut' the circle at some point and unwind it onto the real line.

First we consider the map $\iota : \mathbb{R} \to \mathbb{S}^1$ defined by

$$\iota(t) := [\cos(2\pi t), \sin(2\pi t)]^\top.$$



For any $\eta \in \mathbb{R}$ the restriction $\iota_\eta := \iota \restriction_{(\eta,\eta+1]}$ is a bijective mapping.
Given a probability measure μ on $\mathbb{S}^1$, we can use ι to define a measure $\tilde{\mu}$ on $\mathbb{R}$ by

$$\tilde{\mu}(A) := \mu(\iota(A))$$

whenever A is a Borel set contained in an interval $(\eta, \eta+1]$ or $[\eta, \eta+1)$. Thus $\tilde{\mu}$ is a periodic Borel measure on $\mathbb{R}$, with measure equal to 1 over any period $(\eta, \eta+1]$ or $[\eta, \eta+1)$. Furthermore, $\iota_\eta \# \tilde{\mu} = \mu$.
We define a function $F_\mu : \mathbb{R} \to \mathbb{R}$ by

$$F_\mu(t) := \tilde{\mu}((0,t]) \text{ for } t \in (0,1],$$

and we extend this definition to $\mathbb{R}$ by $F_\mu(t+z) = F_\mu(t) + z$, for all $z \in \mathbb{Z}$. Thus, for $(a,b] \subseteq (\eta, \eta+1]$, we have $F_\mu(b) - F_\mu(a) = \tilde{\mu}((a,b])$.
We define $F_\mu^\eta : [\eta, \eta+1] \to [0,1]$ by

$$F_\mu^\eta(t) := F_\mu(t) - F_\mu(\eta),$$

then F_μ^η is the cumulative distribution function of the probability measure $\tilde{\mu} \restriction_{(\eta,\eta+1]}$ (see Definition 2.3.1). We also define $\left(F_\mu^\eta\right)^{-1} : [0,1] \to [\eta, \eta+1]$ by

$$\left(F_\mu^\eta\right)^{-1}(y) := \inf\{x \in \mathbb{R} | y < F_\mu^\eta(x)\},$$

where we set $\left(F_\mu^\eta\right)^{-1}(1) := \eta + 1$. $\left(F_\mu^\eta\right)^{-1}$ maps into $[\eta, \eta+1]$, because $F_\mu^\eta(x) = 0$ for

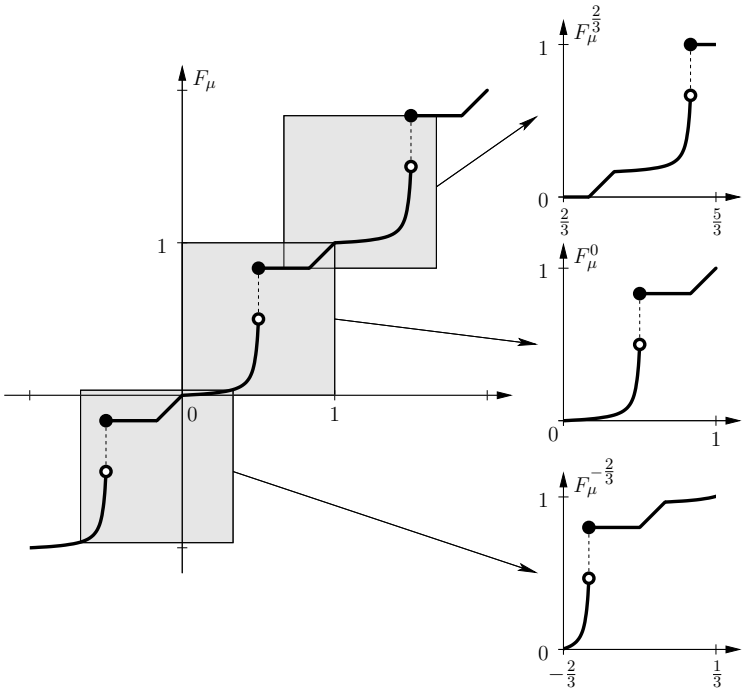


Figure 3.1: Example for a distribution function F_μ of a periodic measure $\tilde{\mu}$. F_μ^η can be seen as a snapshot of F_μ.

$x \leq \eta$ and thus we have $\left(F_\mu^\eta\right)^{-1}(0) \geq \eta$. Also, since $F_\mu^\eta(x) = 1$ for $x \geq \eta + 1$ we have $\left(F_\mu^\eta\right)^{-1}(1^-) \leq \eta + 1$. If we recall Definition 2.3.1 then we notice that $\left(F_\mu^\eta\right)^{-1}(y)$ is the generalized inverse of F_μ^η on $[\eta, \eta+1]$.
If μ and ν are probability measures on $\mathbb{S}^1$, we can consider the measures $\tilde{\mu}$ on $(0,1]$ and $\tilde{\nu}$ on $(\eta, \eta+1]$ and their distribution functions F_μ^0 respectively F_ν^η. Using Theorem 2.3.7 we see that, independent of the strictly convex cost function we use, an optimal transference plan is given by

$$\tilde{\gamma}_\eta = \left(\left(F_\mu^0\right)^{-1} \times \left(F_\nu^\eta\right)^{-1}\right) \# \mathcal{L}^1,$$

because

$$\left|\left(F_\mu^0\right)^{-1}(t) - \left(F_\nu^\eta\right)^{-1}(t)\right| \leq \eta + 2$$

and therefore $I_c(\tilde{\gamma}_\eta) < +\infty$, i.e. $\tilde{\gamma}_\eta$ is finite. This plan induces a transference plan on $\mathbb{S}^1$

by defining $\gamma_\eta := (\iota_0 \times \iota_\eta)\#\tilde{\gamma}_\eta$, because

$$\gamma_\eta(A \times \mathbb{S}^1) = \tilde{\gamma}_\eta(\iota_0^{-1}(A) \times (\eta, \eta+1]) = \iota_0\#\tilde{\mu}(A) = \mu(A),$$
$$\gamma_\eta(\mathbb{S}^1 \times B) = \tilde{\gamma}_\eta((0,1] \times \iota_\eta^{-1}(B)) = \iota_\eta\#\tilde{\nu}(B) = \nu(B),$$

for measurable A and B. Also, if $c(x,y)$ is a cost function on $\mathbb{S}^1$, we get

$$\begin{aligned}\int\limits_{\mathbb{S}^1\times\mathbb{S}^1} c d\gamma_\eta &= \int\limits_{(0,1]\times(\eta,\eta+1]} c(\iota(s),\iota(t)) d\tilde{\gamma}_\eta(s,t) \\ &= \int\limits_{(0,1]\times(\eta,\eta+1]} c\left(\iota\left(\left(F_\mu^0\right)^{-1}(t)\right), \iota\left(\left(F_\nu^\eta\right)^{-1}(t)\right)\right) d\mu(t).\end{aligned}$$

So if we define $\tilde{c}(s,t) := c(\iota(s),\iota(t))$ we get

$$I_c(\gamma_\eta) = \int\limits_{\mathbb{S}^1\times\mathbb{S}^1} c d\gamma_\eta = \int\limits_{(0,1]\times(\eta,\eta+1]} \tilde{c} d\tilde{\gamma}_\eta = I_{\tilde{c}}(\tilde{\gamma}_\eta).$$

Theorem 3.1.1 (Optimal Transport on $\mathbb{S}^1$ for strictly convex cost function)**.** *Let μ, ν be probability measures on $\mathbb{S}^1$ and $c(x,y) = \lambda(d(x,y))$ a convex cost function (i.e. λ is a convex, non-negative and increasing function). Then there exists $\theta \in [-\frac{1}{2}, \frac{1}{2}]$ such that the transference plan γ_θ is optimal. The optimal transport cost is given by*

$$I_c = I_{\tilde{c}}(\tilde{\gamma}_\theta) = \int\limits_0^1 \tilde{c}\left(\left(F_\mu^0\right)^{-1}(t), \left(F_\nu^\theta\right)^{-1}(t)\right) dt. \tag{3.2}$$

If μ is absolutely continuous with respect to σ^1, then $T : \mathbb{S}^1 \to \mathbb{S}^1$ defined by

$$T := \iota_\theta \circ \left(F_\nu^\theta\right)^{-1} \circ F_\mu^0 \circ \iota_0^{-1} \tag{3.3}$$

is an optimal transport plan.

Proof. Let κ be an optimal transference plan (which exists by Theorem 2.1.5), thus $I_c = I_c(\kappa)$.

1. Since every transference plan $\tilde{\gamma}_\eta$ on $\mathbb{R}$ leads to a transference plan γ_η on $\mathbb{S}^1$ and because $I_c(\gamma_\eta) = I_{\tilde{c}}(\tilde{\gamma}_\eta)$, we get

$$I_c(\kappa) = \inf_{\gamma \in \Pi(\mu,\nu)} I_c(\gamma) \leq \inf_{-\frac{1}{2} \leq \alpha \leq \frac{1}{2}} I_{\tilde{c}}(\tilde{\gamma}_\eta).$$

2. Next we show that κ in turn induces a transference plan on $\mathbb{R}$. For $\theta \in [-\frac{1}{2}, \frac{1}{2}]$ we define $\tilde{\kappa}_\theta$ on $R_\theta := (0,1] \times (\theta, \theta+1]$ by setting

$$\tilde{\kappa}_\theta(A) = \kappa((\iota \times \iota)(A))$$

for every Borel set $A \subseteq R_\theta$. This is a transference plan between $\tilde{\mu}$ on $(0,1]$ and $\tilde{\nu}$ on $(\theta, \theta+1]$, because

$$\begin{aligned} \tilde{\kappa}_\theta(A \times (\theta, \theta+1]) &= \kappa(\iota(A) \times \mathbb{S}^1) = \mu(\iota(A)) = \tilde{\mu}(A), \\ \tilde{\kappa}_\theta((0,1] \times B) &= \kappa(\mathbb{S}^1 \times \iota(B)) = \nu(\iota(B)) = \tilde{\nu}(B). \end{aligned}$$

We also see that $(\iota_0 \times \iota_\theta)\#\tilde{\kappa}_\theta = \kappa$.

3. We show that $\tilde{c}(s,t)$ is a convex cost function for s, $t \in \Gamma := \{(s,t) \in \mathbb{R}^2 | |s-t| \leq \frac{1}{2}\}$. For s, $t \in \mathbb{R}$ there exist $n \in \mathbb{N}$ such that $|s-t| \in (n, n+1]$ and either $|s-t| - n \leq \frac{1}{2}$ or $n+1-|s-t| < \frac{1}{2}$. Further, we get

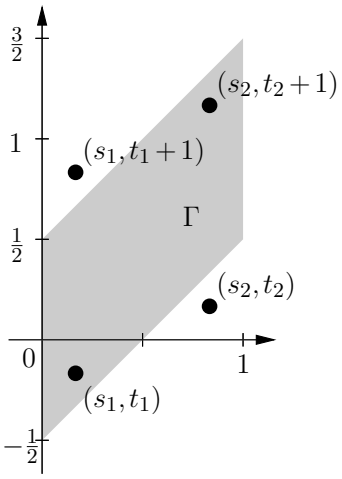
$$\begin{aligned} d(\iota(s), \iota(t)) &= \arccos\left(\begin{pmatrix} \cos(2\pi s) \\ \sin(2\pi s) \end{pmatrix} \cdot \begin{pmatrix} \cos(2\pi t) \\ \sin(2\pi t) \end{pmatrix} \right) \\ &= \arccos(\cos(2\pi s)\cos(2\pi t) + \sin(2\pi s)\sin(2\pi t)) \\ &= 2\pi \min\{|s-t| \bmod 1, 1 - (|s-t| \bmod 1)\}. \end{aligned}$$

So, for s, $t \in \Gamma$ the cost function $\tilde{c}$ is convex, because

$$\tilde{c}(s,t) = \lambda(d(\iota(s), \iota(t))) = \lambda(2\pi|s-t|).$$

4. Next we show that there exists $\theta \in [-\frac{1}{2}, \frac{1}{2}]$ such that $\mathrm{spt}(\tilde{\kappa}_\theta) \subseteq \Gamma$. To prove this, assume no such θ exists. Then there exist pairs

$$(s_1,t_1),(s_2,t_2) \in \bigcup_{\theta\in[-\frac{1}{2},\frac{1}{2}]} \mathrm{spt}(\tilde{\kappa}_\theta) \subseteq (0,1]\times(-\tfrac{1}{2},\tfrac{3}{2}]$$



with $s_1 < s_2$, $t_1 \le t_2$, $|s_1 - t_1| \le \frac{1}{2}$ and $t_2 < s_2 - \frac{1}{2}$. So either $(s_2,t_2) \in \mathrm{spt}(\tilde{\kappa}_\theta)$ or $(s_1,t_1+1) \in \mathrm{spt}(\tilde{\kappa}_\theta)$, but in either case, since $|s_2 - t_2| > \frac{1}{2}$ and $|s_1 - t_1 - 1| > \frac{1}{2}$, we get $(s_2,t_2) \notin \Gamma$ and $(s_1,t_1+1) \notin \Gamma$.
We show that this contradicts the c-cyclical monotonicity of the support of κ and therefore its optimality. We know that

$$\tilde{c}(s_1,t_1) = \lambda(2\pi|s_1 - t_1|) \text{ and } \tilde{c}(s_2,t_2) = \lambda(2\pi|s_2 - t_2 - 1|).$$

Because of

$$\begin{array}{rcccccl} -\frac{1}{2} \le & s_1 - s_2 + \frac{1}{2} & \le & s_1 - t_2 & \le & s_1 - t_1 & \le \frac{1}{2} \\ -\frac{1}{2} \le & s_2 - t_2 - 1 & \le & s_2 - t_1 - 1 & \le & s_2 - s_1 - \frac{1}{2} & \le \frac{1}{2} \end{array}$$

we also get

$$\tilde{c}(s_1,t_2) = \lambda(2\pi|s_1 - t_2|) \text{ and } \tilde{c}(s_2,t_1) = \lambda(2\pi|s_2 - t_1 - 1|).$$

We know that $s_2 - 1 < s_1$ and $t_1 \le t_2$, therefore

$$\tilde{c}(s_1,t_2) + \tilde{c}(s_2,t_1) < \tilde{c}(s_1,t_1) + \tilde{c}(s_2,t_2).$$

Now, since $(s,t) \in \mathrm{spt}(\tilde{\kappa}_\theta)$ implies $(\iota(s),\iota(t)) \in \mathrm{spt}(\kappa)$ this inequality implies, that for $(\iota(s_1),\iota(t_1))$ and $(\iota(s_2),\iota(t_2))$ κ is not c-cyclically monotone. But since an optimal transference plan is always c-cyclically monotone, our initial assumption has to be wrong and there exists $\theta \in [-\frac{1}{2}, \frac{1}{2}]$ such that $\mathrm{spt}(\tilde{\kappa}_\theta) \subseteq \Gamma$.

5. In 2. we saw that $\tilde{\kappa}_\theta$ is a transference plan between $\tilde{\mu}$ on $(0,1]$ and $\tilde{\nu}$ on $(\theta,\theta+1]$. Now, for any convex cost function $\bar{c}$ we know by Theorem 2.3.7 that an optimal transference plan is given by $\tilde{\gamma}_\theta = \left(F_\mu^{-1} \times \left(F_\nu^\theta\right)^{-1}\right) \# \mathcal{L}$ and therefore

$$I_{\bar{c}}(\tilde{\gamma}_\theta) \leq I_{\bar{c}}(\tilde{\kappa}_\theta).$$

So if we extend $\lambda : [0,\pi] \to [0,\infty)$ by defining

$$\lambda(x) = \lambda(\pi) + \frac{\lambda'_-(\pi)}{2}(x-\pi)^2$$

for $x > \pi$ (we recall that λ'_- is the left derivative, see Definition 1.2.3), then $\lambda : [0,+\infty) \to [0,+\infty)$ is still convex, non-negative and increasing. Thus we can define the convex cost function

$$\bar{c}(s,t) = \lambda(2\pi|s-t|)$$

and we have $\bar{c} = \tilde{c}$ on Γ. Hence we get

$$I_{\bar{c}}(\tilde{\gamma}_\theta) = \inf_{\gamma \in \Pi(\tilde{\mu},\tilde{\nu})} I_{\bar{c}}(\gamma) \leq I_{\bar{c}}(\tilde{\kappa}_\theta) = I_{\tilde{c}}(\tilde{\kappa}_\theta) = I_c(\kappa) < +\infty.$$

On the other hand, because of $\bar{c} \geq \tilde{c}$, we have

$$I_{\bar{c}}(\tilde{\gamma}_\theta) \geq I_{\tilde{c}}(\tilde{\gamma}_\theta) = I_c(\gamma_\theta) \geq I_c(\kappa),$$

where the last inequality follows because κ is optimal. Thus we can conclude

$$I_c(\gamma_\theta) = I_c(\kappa)$$

and therefore γ_θ is optimal.

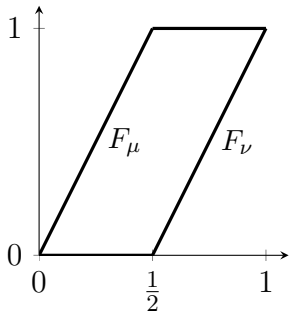
6. Finally, if μ is absolutely continuous with respect to σ^1, then $\tilde{\mu}$ is absolutely continuous with respect to $\mathcal{L}^1$ and by Theorem 2.3.7 the map $\tilde{T} : [0,1] \to [\theta,\theta+1]$ defined by

$$\tilde{T} = \left(F_\nu^\theta\right)^{-1} \circ F_\mu^0$$

is an optimal transport plan between $\tilde{\mu}$ on $(0,1]$ and $\tilde{\nu}$ on $(\theta,\theta+1]$. Thus T, as defined in (3.3), is an optimal transport between μ and ν. □

Example 3.1.2. Consider the quadratic cost function $c(x,y) = \frac{1}{2}d(x,y)^2$ and let μ be the uniform probability measure on the upper half of the circle and ν the uniform probability measure on the lower half of the circle. Then the cumulative distribution functions are given by

$$F_\mu^0(x) = \begin{cases} 2x & \text{if } x \in [0, \frac{1}{2}), \\ 1 & \text{else,} \end{cases}$$

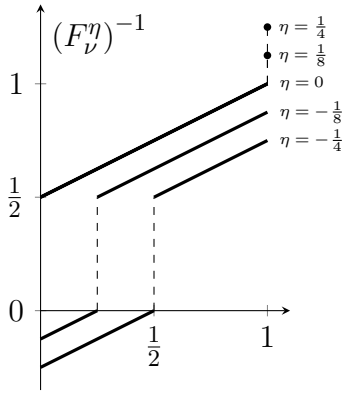


and

$$F_\nu^0(x) = \begin{cases} 0 & \text{if } x \in [0, \frac{1}{2}), \\ 2x-1 & \text{else,} \end{cases}$$

For $\eta \in [-\frac{1}{2}, 0)$ the inverse of F_ν^η is given by

$$(F_\nu^\eta)^{-1}(x) = \begin{cases} \frac{1}{2}x + \eta & \text{if } x \in [0, -2\eta), \\ \frac{1}{2}(x+1) + \eta & \text{if } x \in [-2\eta, 1], \end{cases}$$

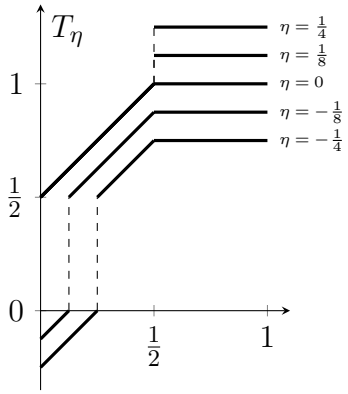


and for $\eta \in [0, \frac{1}{2}]$ the inverse is given by

$$(F_\nu^\eta)^{-1}(x) = \begin{cases} \frac{1}{2}(x+1) & \text{if } x \in [0, 1), \\ 1 + \eta & \text{if } x = 1. \end{cases}$$

Thus for $\eta \in [-\frac{1}{2}, 0)$ the transport map $T_\eta := (F_\nu^\eta)^{-1} \circ F_\mu$ is given by

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



and for $\eta \in [0, \frac{1}{2}]$ the transport map is

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

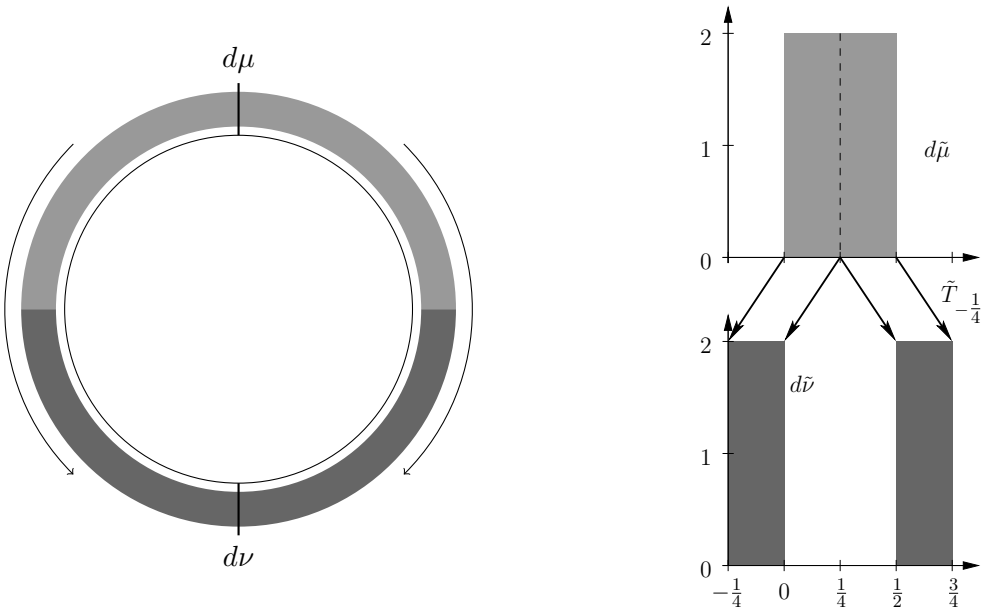
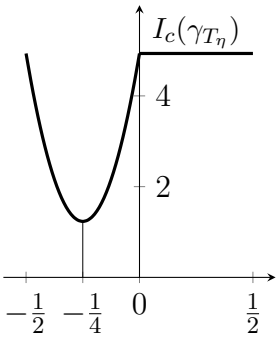


Figure 3.2: Sketch of the solution of Example 3.1.2.

We get

$$I_c(\gamma_{T_\eta}) = \begin{cases} \pi^2(6\eta^2+3\eta+\frac{1}{2}) & \text{if } \eta \in [-\frac{1}{2},0), \\ \frac{1}{2}\pi^2 & \text{if } \eta \in [0,\frac{1}{2}]. \end{cases}$$



This expression becomes minimal for $\eta = -\frac{1}{4}$ and therefore $T_{-\frac{1}{4}}$ is an optimal transport plan. One can think of $T_{-\frac{1}{4}}$ in the following way: First we cut the circle at the bottom ($\iota(-\frac{1}{4}) = (0,-1)^\perp$). Then we move in a counter-clockwise direction from there and transport μ to ν monotonously.

3.2 Calculus on the n-Sphere

Definition 3.2.1 (Differentiability in $\mathbb{R}^n$)**.** A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called **differentiable** at $x \in \mathbb{R}^n$ if there is a linear function $L_x : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$\lim_{y \to x} \frac{\|f(y) - f(x) - L_x(y-x)\|}{\|y - x\|} = 0, \tag{3.4}$$

or, equivalently, if

$$f(x+h) = f(x) + L_x(h) + \mathrm{o}(\|h\|). \tag{3.5}$$

L_x will be denoted by $Df(x)$, the **derivative** of f in x. Furthermore, since $Df(x)$ is linear, there is a matrix $df(x) \in \mathbb{R}^{m \times n}$ such that

$$Df(x)(h) = df(x)h. \tag{3.6}$$

We call this uniquely determined matrix the **Jacobian** of f in x.
If $f : \mathbb{R}^n \to \mathbb{R}$ then $df(x) \in \mathbb{R}^{1 \times n}$ is a vector called the **gradient** of f in x, we denote it by $\nabla f(x)$.

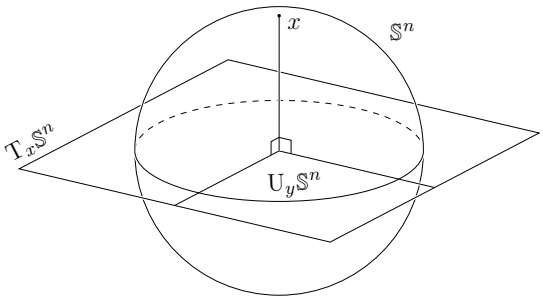
Next we will give a similar definition for functions defined on $\mathbb{S}^n$. To do so we need to introduce the concept of tangent-space.

Definition 3.2.2 (Tangent space)**.** The space of tangent vectors of $\mathbb{S}^n$ at a point $x \in \mathbb{S}^n$ is

$$\mathrm{T}_x\mathbb{S}^n := \{v \in \mathbb{R}^{n+1} | \, \langle v, x \rangle = 0\}. \tag{3.7}$$

Further, we denote the set of unit tangent vectors by

$$\mathrm{U}_x\mathbb{S}^n := \{v \in \mathrm{T}_x\mathbb{S}^n | \, \|v\| = 1\}. \tag{3.8}$$



In the Euclidean case we defined differentiability by the approximation property (3.5). We can use a similar approximation on $\mathbb{S}^n$ to define differentiability for real-valued spherical functions.

Definition 3.2.3 (Differentiability in $\mathbb{S}^n$)**.** The function $\psi : \mathbb{S}^n \to \mathbb{R}$ is called **differentiable** at $x \in \mathbb{S}^n$ if there exists a unique vector $p \in \mathrm{T}_x\mathbb{S}^n$ such that

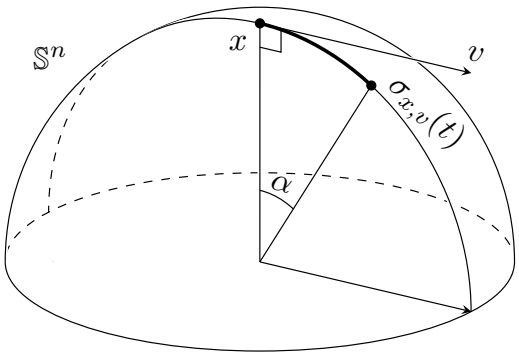
$$\psi(\cos(\alpha)x + \sin(\alpha)v) = \psi(x) + \alpha \langle p, v \rangle + \mathrm{o}(\alpha) \tag{3.9}$$

for all $\alpha > 0$ and $v \in \mathrm{U}_x\mathbb{S}^n$. In this case we will denote p by $\nabla\psi(x)$, the **gradient** of ψ in x.

We should note here that for $x \in \mathbb{S}^n$ and $v \in \mathrm{U}_x\mathbb{S}^n$ the path $\sigma_{x,v} : \mathbb{R} \to \mathbb{S}^n$ defined by

$$\sigma_{x,v}(t) := \cos(t)x + \sin(t)v$$

is geodesic[1], i.e. $d(x, \sigma_{x,v}(t)) = t$ for $t \in [0, \pi]$. This is similar to (3.5) where $x + h$ can be interpreted as a geodesic path, i.e. a straight line from x to $x + h$.This similarity is no coincidence since both $\mathbb{R}^n$ and $\mathbb{S}^n$ are Riemannian manifolds and on Riemannian manifolds differentiability of a function can be expressed by a condition similar to (3.5) resp. (3.9).



We will now give some equivalent conditions for differentiability of functions on $\mathbb{S}^n$.

Proposition 3.2.4. *Let $\psi : \mathbb{S}^n \to \mathbb{R}$ be a function and $\psi^* : \mathbb{R}^{n+1}\backslash\{0\} \to \mathbb{S}^n$ the 0-homogenous extension of ψ from $\mathbb{S}^n$ to $\mathbb{R}^{n+1}\backslash\{0\}$ given by*

$$\psi^*(z) = \psi(\tfrac{z}{\|z\|}).$$

Then ψ^* is differentiable at x if and only if ψ is differentiable at x. Further, in this case $\nabla\psi^*(x) = \nabla\psi(x)$.

Proof.

1. If ψ^* is differentiable at $x \in \mathbb{S}^n$ we fix an arbitrary unit vector $v \in \mathrm{U}_x\mathbb{S}^n$ and define a function $g(\alpha) := \cos(\alpha)x + \sin(\alpha)v$. Using the first order Taylor-approximation of

[1] A path is geodesic if and only if it is locally the shortest connection between curve points.

$\psi^* \circ g$ in 0 we get

$$\begin{aligned}\psi(\cos(\alpha)x + \sin(\alpha)y) = \psi^* \circ g(\alpha) &= \psi(x) + \alpha d(\psi^* \circ g)(0) + \mathrm{o}(\alpha) \\ &= \psi(x) + \alpha \langle \nabla\psi^*(x), v\rangle + \mathrm{o}(\alpha).\end{aligned}$$

For $\nabla\psi^*(x)$ to be the gradient it has to satisfy (3.9). Thus we have to check $\nabla\psi^*(x) \in \mathrm{T}_x\mathbb{S}^n$. Since ψ^* is constant along rays from the origin, the gradient $\nabla\psi^*(x)$ is orthogonal to x and therefore indeed $\nabla\psi(x) = \nabla\psi^*(x)$.

2. Let ψ be differentiable at x. We want to show that

$$\psi^*(z) - \psi(x) - \langle \nabla\psi(x), z - x\rangle = \mathrm{o}(\|z - x\|), \tag{3.10}$$

for $z \to x$. To prove this, we set $z' = \frac{z}{\|z\|}$, $\alpha = d(x, z')$, $v = \frac{z - \langle z,x\rangle x}{\|z - \langle z,x\rangle x\|}$ and get

$$\begin{aligned}\psi^*(z) - \psi(x) = \psi(z') - \psi(x) = \psi(\cos(\alpha)x + \sin(\alpha)v) - \psi(x) &= \alpha \langle p, v\rangle + \mathrm{o}(\alpha), \\ \langle \nabla\psi(x), z - x\rangle &= \|z\| \sin\alpha \langle \nabla\psi(x), v\rangle, \\ \|z - x\| &= \sqrt{\|z\|^2 + 1 - 2\|z\|\cos(\alpha)}.\end{aligned}$$

Combining these equations we get

$$\begin{aligned}&\lim_{z \to x} \frac{\psi^*(z) - \psi(x) - \langle \nabla\psi(x), z - x\rangle}{\|z - x\|} \\ = &\lim_{\|z\| \to 1, \alpha \to 0} \frac{(\alpha - \|z\|\sin\alpha)\langle \nabla\psi(x), v\rangle + \mathrm{o}(\alpha)}{\sqrt{\|z\|^2 + 1 - 2\|z\|\cos(\alpha)}} = 0,\end{aligned}$$

proving the differentiability of ψ^* at x. □

Next we will relax the condition (3.9) to extend the concept of differentiability.

Definition 3.2.5 (Superdifferentiability and supergradient)**.** Let $\varphi : \mathbb{S}^n \to \mathbb{R}$ be a function and fix a point $x \in \mathbb{S}^n$. φ is called **superdifferentiable** at x with **supergradient** $p \in \mathrm{T}_x\mathbb{S}^n$ if for all $v \in \mathrm{U}_x\mathbb{S}^n$ and $\alpha > 0$

$$\varphi(\cos(\alpha)x + \sin(\alpha)v) \leq \varphi(x) + \alpha \langle p, v\rangle + \mathrm{o}(\alpha) \text{ as } \alpha \to 0. \tag{3.11}$$

Further, we denote the **superdifferential** of φ at x by

$$\overline{\partial}_x\varphi := \{p \in \mathrm{T}_x\mathbb{S}^n | p \text{ is supergradient of } \varphi \text{ at } x\}. \tag{3.12}$$

Remark 3.2.6. By reversing the inequalities, we define **subdifferentiability** of a function φ in $x \in \mathbb{S}^n$ with **subgradient** $p \in \mathrm{T}_x\mathbb{S}^n$. Also, we define the **subdifferential** $\underline{\partial}_x\varphi$ as the set of all subgradients of φ in x.

Now, if a function is super- as well as subdifferentiable at x, it is differentiable.

Lemma 3.2.7. *Let $\psi : \mathbb{S}^n \to \mathbb{R}$ be a function and $x \in \mathbb{S}^n$ arbitrary. If ψ is superdifferentiable and subdifferentiable at x, then ψ is differentiable at x and $\overline{\partial}_x\psi = \underline{\partial}_x\psi = \{\nabla\psi(x)\}$.*

Proof. Since ψ is superdifferentiable there is at least one $p \in \overline{\partial}_x\psi$ and because it is subdifferentiable there also is at least one $q \in \underline{\partial}_x\psi$. By combining the inequalities

$$\begin{aligned}\psi(\cos(\alpha)x + \sin(\alpha)v) &\leq \psi(x) + \alpha\langle p, v\rangle + \mathrm{o}(\alpha),\\ \psi(\cos(\alpha)x + \sin(\alpha)v) &\geq \psi(x) + \alpha\langle q, v\rangle + \mathrm{o}(\alpha),\end{aligned}$$

we get

$$0 \leq \alpha\langle p - q, v\rangle + \mathrm{o}(\alpha).$$

Dividing by α and letting $\alpha \to 0$, we get $0 \leq \langle p - q, v\rangle$ and since $v \in \mathrm{U}_x\mathbb{S}^n$ is arbitrary, we necessarily get $p = q$. Since p was arbitrarily chosen from $\overline{\partial}_x\psi$ we get $\overline{\partial}_x\psi = \{q\}$ and analogously $\underline{\partial}_x\psi = \{p\}$, so in fact $\overline{\partial}_x\psi = \underline{\partial}_x\psi = \{p\}$. Thus

$$\psi(\cos(\alpha)x + \sin(\alpha)v) = \psi(x) + \alpha\langle p, v\rangle + \mathrm{o}(\alpha),$$

hence ψ is differentiable at x with gradient $\nabla\psi(x) = p$. □

3.3 Rademacher's Theorem

In this section we will introduce a version of Rademacher's Theorem on the sphere $\mathbb{S}^n$. First, we recall the classic version.

Theorem 3.3.1 (Rademacher's Theorem)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a locally Lipschitz-continuous function. Then f is differentiable $\mathcal{L}^n$-a.e. and $df : \mathbb{R}^n \to \mathbb{R}^{m\times n}$ is a Borel map.*

Idea of the poof. A rigorous proof can be found in [EG92], we will only state the general idea. We can reduce the problem to Lipschitz-continuous functions $f : \mathbb{R}^n \to \mathbb{R}$. Let $v \in \mathbb{S}^{n-1}$ be an arbitrary direction and define

$$\overline{D}_v f(x) := \lim_{k\to\infty} \sup_{\substack{0<|t|<\frac{1}{k}\\ t\in\mathbb{Q}}} \frac{f(x+tv)-f(x)}{t}$$

and

$$\underline{D}_v f(x) := \lim_{k\to\infty} \inf_{\substack{0<|t|<\frac{1}{k}\\ t\in\mathbb{Q}}} \frac{f(x+tv)-f(x)}{t}.$$

Then $\underline{D}_v f$ and $\overline{D}_v f$ are Borel maps and obviously $\underline{D}_v f \le \overline{D}_v f$. We define

$$A_v := \{x \in \mathbb{R}^n | \underline{D}_v f(x) < \overline{D}_v f(x)\}.$$

Now the function $\varphi : \mathbb{R} \to \mathbb{R}$ defined by $\varphi(t) := f(x+tv)$ is Lipschitz-continuous, therefore absolutely continuous, thus differentiable $\mathcal{L}^1$-a.e. Hence $A_v \cap L$ has Hausdorff measure zero for all lines L parallel to v. Using Fubini's Theorem we conclude that $\mathcal{L}^n(A_v) = 0$, thus the directional derivative

$$D_v f(x) = \lim_{t\to 0} \frac{f(x+tv)-f(x)}{t}$$

exists for $\mathcal{L}^n$-a.e. x and $D_v f$ is a Borel map. Thus we can define

$$\operatorname{grad} f(x) := \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x)\right)$$

for $\mathcal{L}^n$-a.e. x. $\operatorname{grad} f$ is also a Borel map.
Next, one shows that $D_v f(x) = v \cdot \operatorname{grad} f(x)$ for $\mathcal{L}^n$ a.e. x. Then, for $v \in \mathbb{S}^{n-1}$, we define

$$F_v := \{x \in \mathbb{R}^n | D_v f \text{ and } \operatorname{grad} f \text{ exist and } D_v f(x) = v \cdot \operatorname{grad} f(x)\}.$$

Thus for a countable dense subset $(v_k)_{k\in\mathbb{N}} \subseteq \mathbb{S}^{n-1}$ we define

$$F := \bigcap_{k\in\mathbb{N}} F_{v_k},$$

and therefore $\mathcal{L}^n(\mathbb{R}^n \backslash F) = 0$. Finally one shows that $\operatorname{grad} f$ satisfies condition (3.4) on F and therefore f is differentiable on F with gradient $\nabla f(x) = \operatorname{grad} f(x)$. Since f is not differentiable outside of F_v we also have that ∇f is a Borel map. □

Definition 1.2.1 also applies to functions $f : \mathbb{S}^n \to \mathbb{R}^m$, since $\mathbb{S}^n \subseteq \mathbb{R}^{n+1}$. However, it is more natural to link Lipschitz-continuity on the sphere to the chord length metric $d(x,y) = \arccos(\langle x, y\rangle)$.

Lemma 3.3.2. *The function $\psi : \mathbb{S}^{n+1} \to \mathbb{R}^m$ is Lipschitz-continuous if and only if there is a constant $B > 0$ such that*

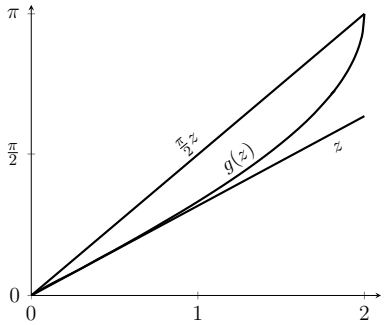
$$\|\psi(x) - \psi(y)\| \leq B\, d(x,y)\ \forall x, y \in \mathbb{S}^n. \tag{3.13}$$

The smallest such constant is given by

$$\operatorname{Lip}_s(f) := \max_{x,y\in\mathbb{S}^n, x\neq y} \frac{\|\psi(x) - \psi(y)\|}{d(x,y)} \leq \operatorname{Lip}(f). \tag{3.14}$$

Proof. We first notice that

$$d(x,y) = \arccos\left(1 - \tfrac{\|x-y\|^2}{2}\right).$$



Thus, since $g(z) := \arccos(1 - \frac{z^2}{2})$ is an increasing convex function on $[0,2]$ with $g(0) = 0$ and $g(2) = \frac{\pi}{2}$ we have $g(z) \leq \frac{\pi}{2} z$.
On the other hand, we have

$$\arccos(1 - \tfrac{z^2}{2})' = \frac{2}{\sqrt{4 - z^2}}, \qquad \text{for } z \in [0,2),$$

hence $g' \geq 1$ and increasing, thus $g(z) \geq z$. We can conclude that

$$\|x-y\| \leq d(x,y) \leq \tfrac{\pi}{2}\|x-y\|$$

and therefore condition (3.13) is equivalent to (1.5). Also

$$\frac{\|\psi(x)-\psi(y)\|}{d(x,y)} \leq \frac{\|\psi(x)-\psi(y)\|}{\|x-y\|} \leq \frac{\pi}{2}\frac{\|\psi(x)-\psi(y)\|}{d(x,y)}$$

and therefore $\mathrm{Lip}_s(f) \leq \mathrm{Lip}(f) \leq \frac{\pi}{2}\mathrm{Lip}_s(f)$. □

Note that since $\mathbb{S}^n$ is compact, a locally Lipschitz-continuous map $\psi : \mathbb{S}^n \to \mathbb{R}^m$ is always Lipschitz-continuous.

To prove a version of Rademacher's theorem on $\mathbb{S}^n$ we will use

Definition 3.3.3 (Hyperspherical coordinates)**.** For the n-sphere $\mathbb{S}^n$ we define the *hypersperical coordinates* $\eta^N : U \subseteq \mathbb{R}^n \to \mathbb{S}^n$ centred in $N := (0,\dots,0,1)$ by

$$\begin{aligned}
\eta_1^N(y) &= \sin(y_1)\\
\eta_2^N(y) &= \cos(y_1)\sin(y_2)\\
&\vdots\\
\eta_n^N(y) &= \cos(y_1)\cos(y_2)\dots\cos(y_{n-1})\sin(y_n)\\
\eta_{n+1}^N(y) &= \cos(y_1)\cos(y_2)\dots\cos(y_{n-1})\cos(y_n)
\end{aligned}$$

where $U := [-\frac{\pi}{2},\frac{\pi}{2}]^{n-1} \times [-\pi,\pi)$.
Furthermore, we define coordinates centered in $x \in \mathbb{S}^n$ by setting $\eta^x(y) = G_x \cdot \eta^N(y)$ where $G_x \in SO(n+1)$ with $G_x \cdot N = x$, thus $\eta^x(0) = x$ (Note that G_x is not unique!).
The inverse map $\zeta^N = (\eta^N)^{-1}$ is given by

$$\begin{aligned}
\zeta_1^N(x) &= \arctan\left(\frac{x_1}{\sqrt{x_{n+1}^2+\dots+x_2^2}}\right)\\
\zeta_2^N(x) &= \arctan\left(\frac{x_2}{\sqrt{x_{n+1}^2+\dots+x_3^2}}\right)\\
&\vdots
\end{aligned}$$

$$\zeta^N_{n-1}(x) = \arctan\left(\frac{x_{n-1}}{\sqrt{x_{n+1}^2 + x_n^2}}\right)$$
$$\zeta^N_n(x) = 2\arctan\left(\frac{x_n}{\sqrt{x_{n+1}^2 + x_n^2} + x_{n+1}}\right)$$

and we note that $\zeta^x(z) = \zeta^N(G_x^\top z)$. Also, η^x as a function from the interior of its domain to $\mathbb{R}^{n+1}$, is $\mathcal{C}^\infty$ in the Euclidean sense. Furthermore, since

$$\zeta^x(\tfrac{z}{\|z\|}) = \zeta^x(z),$$

we can regard ζ^x as a function from $\mathbb{R}^{n+1}\backslash\{0\}$ to $\mathbb{R}^n$ and as such it is $\mathcal{C}^\infty$ in the Euclidean sense as well.

Using these coordinates we can express a function $\psi : \mathbb{S}^n \to \mathbb{R}$ locally around $x \in \mathbb{S}^n$ as a function in $\mathbb{R}^n$ by $\psi \circ \eta_x$. We will now show that $\psi \circ \eta_x$ is differentiable at 0 if and only if ψ is differentiable at x.

Proposition 3.3.4. *Let $\psi : \mathbb{S}^n \to \mathbb{R}$ be a function and $y \in \mathbb{S}^n$ arbitrary. Then ψ is differentiable at $x \in \mathbb{S}^n$ if and only if $\psi \circ \eta^y$ is differentiable at $\zeta^y(x)$. Further, the gradient of ψ at x is determined by*

$$\nabla\psi(x) = \nabla(\psi \circ \eta^y)(\zeta^y(x)) \cdot d(\zeta^y)(x), \tag{3.15}$$

where $d(\zeta^y)$ is the Jacobian of ζ^y when ζ^y is regarded as function from $\mathbb{R}^n\backslash\{0\}$ to $\mathbb{R}^n$.

Proof. If ψ is differentiable at x then, by definition, the extension ψ^* to $\mathbb{R}^{n+1}\backslash\{0\}$ is differentiable at x. Since η^y is differentiable as a function from $\mathbb{R}^n$ to $\mathbb{R}^{n+1}$ we get that $\psi^* \circ \eta^y = \psi \circ \eta^y$ is differentiable at $\zeta^y(x)$.
Next, we assume that $\psi \circ \eta^y$ is differentiable at $\zeta^y(x)$. We consider the function $g(z) = \frac{z}{\|z\|}$. Since the inverse coordinates satisfy $\zeta^y \circ g = \zeta^y$ and are differentiable at x in the Euclidean sense, we can conclude that

$$\psi^* = \psi \circ g = (\psi \circ \eta^y) \circ (\zeta^x \circ g) \tag{$\star$}$$

is differentiable at x and thus, by Proposition 3.2.4 ψ is differentiable at x as well. Finally,

by using the chain rule on ($\star$), we get

$$\nabla\psi^*(x) = \nabla(\psi \circ \eta^y)(\zeta^y(x)) \cdot d(\zeta^y)(x),$$

thus we are done, since $\nabla\psi(x) = \nabla\psi^*(x)$ for $x \in \mathbb{S}^n$. □

Theorem 3.3.5 (Rademacher's Theorem for $\mathbb{S}^n$)**.** *Let $\psi : \mathbb{S}^n \to \mathbb{R}$ be Lipschitz-continuous, then ψ is differentiable σ^n-a.e. and $\nabla\psi$ is a Borel map.*

Proof. Let $x \in \mathbb{S}^n$, $\epsilon > 0$ and choose coordinates η^x. We will show that ψ is differentiable σ^n-a.e. in a neighborhood $B_\epsilon(x) = \{y \in \mathbb{S}^n | d(x,y) < \epsilon\}$. There exists $\delta > 0$ such that $\eta^x(B_\delta(0)) \supseteq B_\epsilon(x)$. Since $\psi : \mathbb{S}^n \to \mathbb{R}$ is Lipschitz-continuous, $\psi \circ \eta^x$ is Lipschitz-continuous from $\mathbb{R}^n$ to $\mathbb{R}$. Now Rademacher's Theorem guarantees that there is a Lebesgue-null set $Z \subseteq B_\delta(0)$ such that $\psi \circ \eta^x$ is differentiable on $B_\delta(0) \backslash Z$ and that $\nabla(\psi \circ \eta^x)$ is a Borel map. Thus by Proposition 3.3.4, ψ is differentiable on $F_x := B_\epsilon(x) \backslash \eta^x(Z)$ and therefore σ^n-a.e. since

$$\sigma^n(\eta^x(Z)) = \int\limits_{\eta^x(Z)} d\sigma^n = \int\limits_Z \sqrt{\det((d\eta^x(y))^\top d\eta^x(y))} d\mathcal{L}^n(y) = 0.$$

Further, using (3.15), we see that the gradient $\nabla\psi$ is a Borel map on F_x. Thus, since $\mathbb{S}^n$ is compact, there are $x_1, x_2, \ldots, x_k$ such that $\bigcup_{i=1}^k B_\epsilon(x_i) = \mathbb{S}^n$ and hence ψ is differentiable on $F := \bigcup_{i=1}^k F_{x_i}$ and $\sigma^n(\mathbb{S}^n \backslash F) = 0$. Therefore ψ is differentiable σ^n-a.e. and $\nabla\psi$ is a Borel map. □

3.4 Optimal Transport on the n-Sphere

In this section we will prove that there is a σ^n-a.e. unique solution to the MP on the n-sphere, if the cost function is $c(x,y) = \lambda(d(x,y))$ and μ is absolutely continuous with respect to σ^n.

McCann showed in [McC01] that the MKP can be solved on Riemannian manifolds for strictly convex cost function. Our main Theorem 3.4.9 can be seen as a corollary to Theorem 13 in [McC01]. To prove Theorem 3.4.9, we will follow mostly the same strategy as McCann, but we will restrict ourselves to the n-sphere.

First we show that a c-concave function is Lipschitz-continuous. To do so, we need some properties of the chord length metric $d(x,y) = \arccos(\langle x, y\rangle)$.

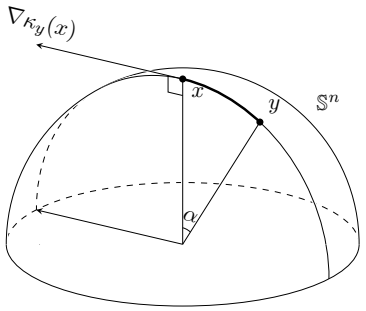
Lemma 3.4.1. *Let $y \in \mathbb{S}^n$ be arbitrary and $d(x,y) = \arccos(\langle x, y\rangle)$. Define $\kappa_y : \mathbb{S}^n \to [0,\pi]$ by $\kappa_y(x) := d(x,y)$. Then κ_y has the following properties:*

(i) κ_y is Lipschitz-continuous with Lipschitz-constant $\mathrm{Lip}_s(\kappa_y) \leq 1$.

(ii) κ_y is differentiable for $x \neq \pm y$ with gradient

$$\nabla \kappa_y(x) = \tfrac{1}{\sin\alpha}(\cos(\alpha)x - y) \tag{3.16}$$

where $\alpha = \kappa_y(x) = d(x,y)$.



(iii) κ_y is superdifferentiable at $x = -y$ with superdifferential

$$\overline{\partial}_{-y}\kappa_y = \mathrm{U}_y\mathbb{S}^n. \tag{3.17}$$

Proof.

ad (i): Since d is a metric we get

$$d(x,y) \leq d(x,z) + d(y,z),$$
$$d(z,y) \leq d(z,x) + d(y,x)$$

and therefore

$$\kappa_y(x) - \kappa_y(z) \leq d(x,z),$$
$$\kappa_y(z) - \kappa_y(x) \leq d(x,z).$$

So we obtain $\mathrm{Lip}_s(\kappa_y) \leq 1$.

ad (ii): For $x \neq \pm y$, $\kappa_y(x) = \arccos(\langle x, y\rangle)$ is obviously differentiable at x. We extend κ_y to $\mathbb{R}^{n+1}\backslash\{0\}$ by $\kappa_y^*(x) := \kappa_y(\frac{x}{\|x\|})$. Using the chain rule we can compute the differential

of κ_y^* at x

$$\nabla\kappa_y^*(x) = \frac{-1}{\sqrt{1-\left\langle \frac{x}{\|x\|}, y\right\rangle^2}}\left(\frac{y}{\|x\|} - \frac{\langle x,y\rangle}{\|x\|^3}x\right).$$

For $x \in \mathbb{S}^n$ and $\alpha = d(x,y)$ we get

$$\nabla\kappa_y(x) = \nabla\kappa_y^*(x) = \tfrac{1}{\sin\alpha}(\cos(\alpha)x - y).$$

ad (iii): Let $x = -y$ and choose $z \neq \pm y$. We have

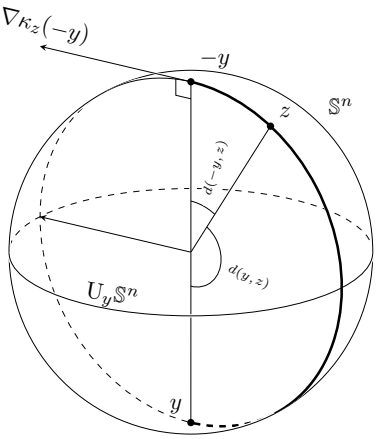
$$\pi = d(-y,y) = d(-y,z) + d(z,y).$$

By (ii) κ_z is differentiable at $-y$, thus for $\beta > 0$ and $w \in \mathrm{U}_{-y}\mathbb{S}^n$ we have

$$\begin{aligned} &d(z, \cos(\beta)(-y) + \sin(\beta)w) \\ &= \kappa_z(\cos(\beta)(-y) + \sin(\beta)w) \\ &= \kappa_z(-y) + \beta\,\langle \nabla\kappa_z(-y), w\rangle + \mathrm{o}(\beta). \end{aligned}$$

Using the triangle inequality we conclude

$$\begin{aligned} &\kappa_y(\cos(\beta)(-y) + \sin(\beta)w) \\ &\quad\leq d(y,z) + d(z, \cos(\beta)(-y) + \sin(\beta)w) \\ &\quad= \underbrace{d(y,z) + d(-y,z)}_{d(-y,y)} + \beta\,\langle \nabla\kappa_z(-y), w\rangle + \mathrm{o}(\beta). \end{aligned}$$



Therefore $\nabla\kappa_z(-y)$ is a supergradient of κ_y in $-y$. Since $\nabla\kappa_z(-y) \in \mathrm{U}_{-y}\mathbb{S}^n$ and because

$$\nabla\kappa_z(-y) = \frac{\cos(d(-y,z))(-y) - z}{\sin(d(-y,z))} = \frac{\cos(d(y,z))y - z}{\sin(d(y,z))},$$

we get that $\nabla\kappa_z(-y)$ points in the direction of $y - z$ when projecting $y - z$ onto $\mathrm{U}_y\mathbb{S}^n$. Hence $\underline{\partial}_{-y}\kappa_y = \mathrm{U}_y\mathbb{S}^n$. □

For our main Theorem 3.4.9 we require the strictly convex cost function c to satisfy some

further properties.

Proposition 3.4.2. *Let $\lambda : [0,\pi] \to [0,+\infty)$ and $d(x,y) = \arccos(\langle x,y\rangle)$. We will call a cost function $c(x,y) = \lambda(d(x,y))$* ***strictly convex**** *if λ satisfies the following properties:*

(i) λ is strictly convex on $[0,\pi]$.

(ii) λ is differentiable on $(0,\pi)$.

(iii) λ satisfies

$$\lim_{t\to 0^+} \frac{\lambda(t)}{t} = 0,$$

and

$$\lambda'_-(\pi) = \lim_{t\to 0^+} \frac{\lambda(\pi)-\lambda(\pi-t)}{t} < +\infty.$$

Given a strictly convex cost function c and arbitrary $y \in \mathbb{S}^n$, the function $\tau : \mathbb{S}^n \to [0,+\infty)$ defined by $\tau(x) = c(x,y)$ is Lipschitz-continuous and differentiable for $x \neq -y$. The gradient of τ in $x \neq \pm y$ is given by*

$$\nabla(\lambda\circ\tau)(x) = \tfrac{\lambda'(\alpha)}{\sin\alpha}(\cos(\alpha)x - y), \tag{3.18}$$

where $\alpha = d(x,y)$ and $\nabla(\lambda\circ\tau)(y) = 0$. Furthermore, for $x = -y$, τ is superdifferentiable with superdifferential

$$\overline{\partial}_{-y}(\lambda\circ\tau) = \lambda'_-(\pi)\mathrm{U}_y\mathbb{S}^n.$$

Proof. Since λ is differentiable on $(0,\pi)$ and $d(x,y)$ is bounded by $0 < d(x,y) < \pi$ for $x \neq \pm y$, we see by Lemma 3.4.1 that τ is differentiable for $x \neq \pm x$. By using the chain rule we get (3.18).
For $x = y$ we get $\tau(y) = 0$ and for $w \in \mathrm{U}_y\mathbb{S}^n$ and $\beta > 0$ we have

$$\tau(\cos(\beta)y + \sin(\beta)w) = \lambda(\arccos(\langle\cos(\beta)y + \sin(\beta)w, y\rangle)) = \lambda(\beta),$$

thus

$$\lim_{\beta\to 0^+} \frac{\tau(\cos(\beta)y+\sin(\beta)w)-\tau(y)}{\beta} = \lim_{\beta\to 0^+}\frac{\lambda(\beta)}{\beta} = 0.$$

Therefore τ is differentiable at $x = y$ with gradient $\nabla\tau(y) = 0$.
Finally, for $x = -y$, we have $\tau(-y) = \lambda(d(-y,y)) = \lambda(\pi)$. Since λ is convex and differentiable in $(0,1)$, we have

$$\lim_{\epsilon\to 0^+} f'(\pi-\epsilon) = f'_-(\pi) < \infty,$$

thus for $\epsilon > 0$ we have

$$\lambda(\pi-\epsilon) \le \lambda(\pi) - \epsilon\lambda'_-(\pi) + \mathrm{o}(\epsilon).$$

Now, let $v \in \mathrm{U}_y\mathbb{S}^n$. Then by Lemma 3.4.1 $-v$ is a superdifferential of $x \mapsto d(x,y)$ at $-y$. Hence, for $w \in \mathrm{U}_{-y}\mathbb{S}^n$ and $\beta > 0$ we set $\epsilon = \beta\langle v, w\rangle + \mathrm{o}(\beta)$ and, since λ is increasing, we conclude

$$\lambda(\tau(-\cos(\beta)y + \sin(\beta)w)) \le \lambda(\tau(-y) - \beta\langle v,w\rangle + \mathrm{o}(\beta)) \le \lambda(\pi) - \beta\lambda'_-(\pi)\langle v,w\rangle + \mathrm{o}(\beta).$$

Therefore $-\lambda'_-(\pi)v$ is a supergradient of τ at $-y$. □

Example 3.4.3. The quadratic cost function $c(x,y) = \frac{1}{2}d(x,y)^2$ is strictly convex* since the function $\lambda(z) = \frac{z^2}{2}$ satisfies all required properties of Proposition 3.4.2.

Proposition 3.4.4 (c-concave functions are Lipschitz-continuous)**.**
Let $c(x,y) = \lambda(d(x,y))$ be a cost function, with $\lambda : [0,\pi) \to [0,+\infty)$ strictly convex and increasing. Let $\psi : \mathbb{S}^n \to \mathbb{R}\cup\{-\infty\}$ be a c-concave function (i.e. $\psi^{cc} = \psi$ and $\{\psi = -\infty\} \neq \mathbb{S}^n$). Then ψ and its c-transform ψ^c are both Lipschitz-continuous and the Lipschitz-constant of ψ is bounded by $B := \sup_{y\in\mathbb{S}^n} \mathrm{Lip}_s(\tau_y) < +\infty$, where $\tau_y(x) = c(x,y)$.

Proof. Since the distance function $d(x,y)$ is bounded from above by π, the cost function is bounded by $0 \le c(x,y) \le \lambda(\pi)$. Further, because ψ is c-concave, it follows that $\psi(x) = \inf_{y\in\mathbb{S}^n} c(x,y) - \psi^c(y)$. Thus ψ^c is bounded from above, because otherwise $\psi \equiv -\infty$, and therefore ψ is bounded from below. Thus, since $\psi < +\infty$ and $\mathbb{S}^n$ is compact, we see that ψ is bounded.
Fix $z \in \mathbb{S}^n$, then $\forall\epsilon > 0$ there exists y such that

$$\psi(z) + \epsilon \ge c(z,y) - \psi^c(y).$$

Furthermore, by definition, we have

$$\psi(x) \leq c(x,y) - \psi^c(y).$$

Combining these inequalities we get

$$\psi(x) - \psi(z) \leq c(x,y) - c(z,y) + \epsilon.$$

The map $\tau_y : x \mapsto c(x,y)$ is Lipschitz-continuous because λ is convex and therefore Lipschitz-continuous and the map $x \mapsto d(x,y)$ is Lipschitz-continuous by Lemma 3.4.1. Thus we get

$$\psi(x) - \psi(z) \leq c(x,y) - c(z,y) + \epsilon \leq \mathrm{Lip}_s(\tau_y) d(x,z) + \epsilon.$$

Thus ψ is Lipschitz-continuous and the Lipschitz-constant of ψ satisfies

$$\mathrm{Lip}_s(\psi) \leq \sup_{y \in \mathbb{S}^n} \mathrm{Lip}_s(\tau_y) = B,$$

which is finite because $\mathrm{Lip}_s(\tau_y) < +\infty$ and $\mathbb{S}^n$ is compact. □

We recall the following result.

Proposition 3.4.5 (The dual MKP has maximizers.)**.** *Let $c(x,y) = \lambda(d(x,y))$ be a cost function, with $\lambda : [0,\pi] \to [0,\infty)$ strictly convex and increasing. Let μ and ν be probability measures on $\mathbb{S}^n$, then there exists a c-concave function ψ which satisfies*

$$J(\psi, \psi^c) = \sup_{(u,v) \in \Phi_c} J(u,v),$$

where

$$J(u,v) = \int_{\mathbb{S}^n} u d\mu + \int_{\mathbb{S}^n} v d\nu$$

and

$$\Phi_c = \{(u,v) \in \mathrm{L}^1(\mathbb{S}^n, \mu) \times \mathrm{L}^1(\mathbb{S}^n, \nu) | u(x) + v(y) \leq c(x,y)\}.$$

Proof. This is clear from Corollary 2.2.9, since c is continuous and bounded. □

Remark 3.4.6. Proposition 3.4.5 could also be proved directly. First, we may restrict ourselves to tight pairs $(u,v) \in \Phi_c$ to maximize $J(u,v)$, thus we may assume u to be c-concave and $v = u^c$. Then let u_n be a sequence of c-concave functions such that $\lim_{n\to\infty} J(u_n, u_n^c)$ is maximal. Since c-concave functions prove to be Lipschitz-continuous, one can apply the **Arzelà-Ascoli Theorem** to extract a converging subsequence with Lipschitz-continuous limits $(u^*, v^*) \in \Phi_c$. Using dominated convergence

$$\lim_{n\to\infty} J(u_n, u_n^c) = J(u^*, v^*).$$

Finally, one puts $\psi := (v^*)^c$. So ψ is c-concave and the pair (ψ, ψ^c) maximizes the functional J. A rigorous proof using this method can be found in [McC01] (Proposition 3).

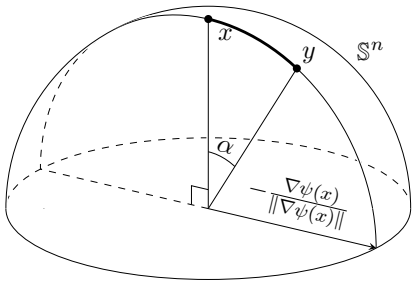
By Proposition 3.4.5 there is always a solution ψ to the dual MKP. This solution is c-concave and therefore, by Proposition 3.4.4, Lipschitz-continuous. Rademacher's Theorem 3.3.5 shows that the maximizing function is differentiable σ^n-a.e. with a gradient $\nabla\psi$ that is a Borel map. We will use this gradient to define a transport plan.

Proposition 3.4.7. *Let $c(x,y) = \lambda(d(x,y))$ be a strictly convex* cost function (as defined in Proposition 3.4.2) with $\lambda : [0,\pi] \to [0,+\infty)$. Also, let $\psi : \mathbb{S}^n \to \mathbb{R}$ be c-concave and $x \in \mathbb{S}^n$ arbitrary. Let ψ be differentiable at x and define*

$$\alpha := (\lambda^*)'(\|\nabla\psi(x)\|),$$

where λ^* is the conjugate function of λ (which is differentiable since λ is strictly convex). Then $y \in \mathbb{S}^n$ satisfies $\psi(x) + \psi^c(y) = c(x,y)$, if and only if

$$y = \cos(\alpha)x - \sin(\alpha)\frac{\nabla\psi(x)}{\|\nabla\psi(x)\|}. \tag{3.19}$$



Note that if $\nabla\psi(x) = 0$ then $\alpha = 0$. In this case (3.19) should be read as $y = x$.

Proof. Let ψ be differentiable at $x \in \mathbb{S}^n$ and let $y \in \mathbb{S}^n$ be such that $\psi(x) + \psi^c(y) = c(x,y)$. By definition of the c-transform we get

$$0 = c(x,y) - \psi(x) - \psi^c(y) \le c(z,y) - \psi(z) - \psi^c(y).$$

Now let $t > 0$ and $v \in \mathrm{U}_x\mathbb{S}^n$ such that $z = \cos(t)x + \sin(t)v$. Since ψ is differentiable at x, we obtain

$$\psi(z) = \psi(\cos(t)x + \sin(t)v) = \psi(x) + t\,\langle \nabla\psi(x), v\rangle + \mathrm{o}(t).$$

Using these results we get

$$c(z,y) \geq c(x,y) - \psi(x) + \psi(z) = c(x,y) + t\,\langle \nabla\psi(x), v\rangle + \mathrm{o}(t),$$

and therefore $\nabla\psi(x)$ is a subgradient at x of the function $\tau : x \mapsto c(x,y)$. Using Proposition 3.4.2 we see that τ is always superdifferentiable and therefore τ is actually differentiable. For $x \neq y$ the gradient of τ at x is

$$\nabla\tau(x) = \frac{\lambda'(\beta)}{\sin\beta}(\cos(\beta)x - y), \tag{$\star$}$$

where $\beta = d(x,y)$. For $x = y$ we have $\nabla\tau(x) = 0$. Since $\nabla\tau(x) = \nabla\psi(x)$ we get

$$\|\nabla\psi(x)\| = \|\nabla\tau(x)\| = \lambda'(\beta) = \lambda'(d(x,y)),$$

or $\|\nabla\psi(x)\| = 0$ if $x = y$. Now, since $(\lambda^*)'(\lambda'(\beta)) = \beta$ for $\beta \in (0,\pi)$ and because $0 \in \underline{\partial}_0\lambda$ implies $(\lambda^*)'(0) = 0$, we have

$$\alpha = (\lambda^*)'(\|\nabla\psi(x)\|) = d(x,y) = \beta.$$

Thus, by using this equation in ($\star$) and the fact that $\lambda'((\lambda^*)'(t)) = t$, we conclude

$$\nabla\psi(x) = \frac{\|\nabla\psi(x)\|}{\sin\alpha}\,(\cos\alpha x) - y)$$

which proves equation (3.19).
We just showed that if there exists $y \in \mathbb{S}^n$ with $\psi(x) + \psi^c(y) = c(x,y)$, then it is uniquely defined by (3.19). We still need to prove that such a y always exists. Since ψ is c-concave we know that

$$\psi(x) = \psi^{cc}(x) = \inf_{y\in\mathbb{S}^n} c(x,y) - \psi^c(y).$$

Now the function $y \mapsto c(x,y) - \psi^c(y)$ is continuous (since $y \mapsto c(x,y)$ and ψ^c are Lipschitz-

continuous) and $\mathbb{S}^n$ is compact, thus there exists $y_0 \in \mathbb{S}^n$ such that

$$\psi(x) = c(x, y_0) - \psi^c(x).$$

Thus this proof is completed. □

Theorem 3.4.8. *Let $c(x,y) = \lambda(d(x,y))$ be a strictly convex* cost function as defined in Proposition 3.4.2. Let μ be a probability measure on $\mathbb{S}^n$ absolutely continuous with respect to σ^n and let $\psi : \mathbb{S}^n \to \mathbb{R}$ be c-concave. We define $T : \mathbb{S}^n \to \mathbb{S}^n$ by*

$$T(x) := \begin{cases} \cos(\alpha)x - \sin(\alpha)\frac{\nabla\psi(x)}{\|\nabla\psi(x)\|} & \text{if } \psi \text{ is differentiable at } x \text{ and } \nabla\psi(x) \neq 0 \\ x & \text{else,} \end{cases} \tag{3.20}$$

where $\alpha = (\lambda^*)'(\|\nabla\psi(x)\|)$. Then T is a Borel map and $\nu := T\#\mu$ is a probability measure. Further, T is the solution to the MKP for μ and ν, i.e.

$$I_c(\gamma_T) = \inf_{\gamma \in \Pi(\mu,\nu)} I_c(\gamma),$$

where $\gamma_T = (\mathrm{id} \times T)\#\mu$. Moreover, if S is another optimal transport map for μ and ν then $\mu[S \neq T] = 0$, i.e. T equals S μ-a.e.

Proof. Because of Proposition 3.4.4 ψ is Lipschitz-continuous and therefore, according to Theorem 3.3.5, ψ is μ-a.e. differentiable and the gradient $\nabla\psi$ is a Borel map. So $T(x)$ is defined μ-a.e. and a Borel map.
Because of Theorem 3.4.8, we have $\psi(x) + \psi^c(T(x)) = c(x, T(x))$ σ^n-a.e. Since μ is absolutely continuous with respect to σ^n we get

$$\begin{aligned} J(\psi, \psi^c) &= \int_{\mathbb{S}^n} \psi d\mu + \int_{\mathbb{S}^n} \psi^c d\nu = \int_{\mathbb{S}^n} \psi(x) + \psi^c(T(x)) d\mu(x) \\ &= \int_{\mathbb{S}^n \times \mathbb{S}^n} c(x, T(x)) d\mu(x) = \int_{\mathbb{S}^n \times \mathbb{S}^n} c d\gamma_T = I_c(\gamma_T). \end{aligned}$$

Hence γ_T is optimal in the MKP.
If $S : \mathbb{S}^n \to \mathbb{S}^n$ is another optimal transport plan for μ and ν, i.e. $S\#\mu = \nu$, then necessarily $I_c(\gamma_S) = J(\psi, \psi^c)$. Therefore $\psi(x) + \psi^c(S(x)) = c(x, S(x))$ μ-a.e. and because of Theorem 3.4.7 it follows that $S = T$ μ-a.e. □

We can now prove the main result:

Theorem 3.4.9 (Optimal Transport on the n-sphere for a strictly convex* cost function)**.** *Let $c(x,y) = \lambda(d(x,y))$ be a strictly convex* cost function, μ and ν probability measures on $\mathbb{S}^n$ with μ absolutely continuous with respect to σ^n. Then there exists a c-concave function $\psi : \mathbb{S}^n \to \mathbb{R}$ such that the map $T : \mathbb{S}^n \to \mathbb{S}^n$, defined in Theorem 3.4.8, solves the Optimal Transport Problem, i.e. $T\#\mu = \nu$ and*

$$I_c(\gamma_T) = \inf_{\gamma \in \Pi(\mu,\nu)} \int c(x,y) d\gamma(x,y). \tag{3.21}$$

Furthermore, T is μ-a.e. uniquely determined.

Proof.

1. Because of Proposition 3.4.5 there exists a c-concave ψ, such that

$$J(\psi, \psi^c) = \sup_{(u,v)\in\Phi_c} J(u,v).$$

   Thus using Theorem 3.4.8 we can μ-a.e. define

$$T(x) = \cos(\alpha)x - \sin(\alpha)\frac{\nabla\psi(x)}{\|\nabla\psi(x)\|},$$

   where $\alpha = (\lambda^*)'(\|\nabla\psi(x)\|)$.

2. Next we need to show that $T\#\mu = \nu$, or equivalently

$$\int_{\mathbb{S}^n} f \circ T d\mu = \int_{\mathbb{S}^n} f d\nu,$$

   for all $f \in \mathcal{C}(\mathbb{S}^n)$. We fix an arbitrary $f \in \mathcal{C}(\mathbb{S}^n)$ and define

$$\begin{aligned} \tau_\epsilon(y) &:= \psi^c(y) + \epsilon f(y), \\ \psi_\epsilon(x) &:= \tau_\epsilon^c(x) = \inf_{y\in\mathbb{S}^n} c(x,y) - \psi^c(y) - \epsilon f(y). \end{aligned}$$

   ψ_ϵ is c-concave, because $\psi_\epsilon^{cc} = \tau_\epsilon^{ccc} = \tau_\epsilon^c = \psi_\epsilon$.
   We fix $x_0 \in \mathbb{S}^n$ such that $\psi = \psi_0$ is differentiable. Due to Proposition 3.4.7

$$\begin{aligned} c(x_0, T(x_0)) &= \psi_0(x) + \psi_0^c(T(x_0)) = \psi_0(x_0) + \tau_0(T(x_0)), \\ \psi_0(x_0) &= c(x_0, T(x_0)) - \tau_0(T(x_0)). \end{aligned}$$

For small values of ϵ there exists $y_\epsilon \in \mathbb{S}^n$ such that

$$\psi_\epsilon(x_0) = c(x_0, y_\epsilon) - \psi^c(y_\epsilon) - \epsilon f(y_\epsilon). \tag{$\star$}$$

Proposition 3.4.7 also states that $T(x_0)$ is an unique solution and therefore

$$\lim_{\epsilon \to 0^+} y_\epsilon = T(x_0)$$

or equivalently

$$y_\epsilon = T(x_0) + \mathrm{o}(1) \text{ as } \epsilon \to 0^+.$$

Since f is continuous we get

$$\epsilon f(y_\epsilon) = \epsilon f(T(x_0) + \mathrm{o}(1)) = \epsilon f(T(x_0)) + \mathrm{o}(\epsilon)$$

where

$$\tfrac{\mathrm{o}(\epsilon)}{\epsilon} \leq 2 \sup_{x \in \mathbb{S}^n} |f(x)| \leq +\infty.$$

Using this and ($\star$) along with

$$\psi_0(x_0) \leq c(x_0, y_\epsilon) - \psi^c(y_\epsilon),$$

we get a lower bound for ψ_ϵ:

$$\psi_0(x_0) - \epsilon f(T(x_0)) + \mathrm{o}(\epsilon) \leq c(x_0, y_\epsilon) - \psi^c(y_\epsilon) - \epsilon f(y_\epsilon) \leq \psi_\epsilon(x_0).$$

Further, by definition of ψ_ϵ

$$\psi_\epsilon(x_0) \leq c(x_0, y) - \psi^c(y) - \epsilon f(y) \leq \psi_0(x_0) - \epsilon f(T(x_0))$$

and by setting $y = T(x_0)$, we get an upper bound:

$$\psi_\epsilon(x_0) \leq \psi_0(x_0) - \epsilon f(T(x_0)).$$

Therefore

$$\psi_\epsilon(x_0) = \psi_0(x_0) - \epsilon f(T(x_0)) + o(\epsilon).$$

Proposition 3.4.5 also states that $J(\psi_\epsilon, \psi_\epsilon^c)$ is maximal for $\epsilon = 0$ and therefore we get

$$\begin{aligned} 0 &= \lim_{\epsilon\to 0} \frac{J(\psi_\epsilon, \psi_\epsilon^c) - J(\psi_0, \psi_0^c)}{\epsilon} \\ &= \lim_{\epsilon\to 0} \int_{\mathbb{S}^n} \frac{\psi_\epsilon(x) - \psi_0(x)}{\epsilon} d\mu(x) + \int_{\mathbb{S}^n} \frac{\psi_\epsilon^c(y) - \psi_0^c(y)}{\epsilon} d\nu(y) \\ &= \int_{\mathbb{S}^n} -f(T(x)) d\mu(x) + \int_{\mathbb{S}^n} f(y) d\nu(y) + \lim_{\epsilon\to 0} \int_{S^n} \frac{o(\epsilon)}{\epsilon} d\mu. \end{aligned}$$

Since $\frac{o(\epsilon)}{\epsilon}$ is bounded we can use the Theorem of dominated convergence to see that the last integral vanishes. Therefore, since $f \in \mathcal{C}(\mathbb{S}^n)$ was arbitrary, we just proved

$$\int_{\mathbb{S}^n} f \circ T d\mu = \int_{\mathbb{S}^n} f d\nu,$$

i.e. $T\#\mu = \nu$.

3. In order to complete the proof we need to show the uniqueness of T μ-a.e. So, assume that S is another optimal transport plan, i.e. $S\#\mu = \nu$ and $I_c(\gamma_S) = I_c(\gamma_T)$. Then by Theorem 3.4.8 $S = T$ μ-a.e. and we are done. □

Remark 3.4.10. In the classical setting of optimal transport in $\mathbb{R}^n$ with quadratic cost function $c(x,y) = \|x-y\|^2$ the gradient $\nabla\psi$ of a c-concave maximizer $\psi : \mathbb{R}^n \to \mathbb{R}$ of the dual MKP induces an optimal transport plan. In this case, $\nabla\psi$ is more generally known as **Brenier map**. The transport plan T in Theorem 3.4.9 may be seen as a spherical Brenier map.

Corollary 3.4.11 (Inverse Transport)**.** *Let $c(x,y) = \lambda(d(x,y))$ be a strictly convex* cost function. Let μ and ν be probability measures on $\mathbb{S}^n$ with μ and ν absolutely continuous with respect to σ^n. According to Theorem 3.4.9, there exists a c-concave ψ which determines a transport map T. Also, since ν is absolutely continuous with respect to σ^n, ψ^c defines an optimal transport map $T^\dagger$ for the inverse problem, i.e. $T^\dagger\#\nu = \mu$*

and

$$I_c(\gamma_{T^\dagger}) = \sup_{\gamma \in \Pi(\nu,\mu)} \int c(x,y) d\gamma(x,y).$$

Then $T^\dagger$ is inverse to T μ-a.e., i.e. $T \circ T^\dagger = T^\dagger \circ T = \mathrm{id}_{\mathbb{S}^n}$ μ-a.e.

Proof. Theorem 3.4.9 guarantees the existence of T and $T^\dagger$, so we only need to show that $T^\dagger$ is inverse to T μ-a.e. Let

$$U_\psi := \{x \in \mathbb{S}^n | \psi \text{ is differentiable at } x\},$$
$$U_{\psi^c} := \{y \in \mathbb{S}^n | \psi^c \text{ is differentiable at } y\}.$$

Since ψ as well as ψ^c are c-concave they are also Lipschitz-continuous. Thus, since μ and ν are absolutely continuous with respect to σ^n, Rademacher's Theorem 3.3.5 implies that $\mu(U_\psi) = \nu(U_{\psi^c}) = 1$. We define $V := U_\psi \cap T^{-1}(U_{\psi^c})$ to see

$$0 \leq \mu(\mathbb{S}^n \backslash V) \leq \mu\left(\mathbb{S}^n \backslash U_\psi\right) + \mu\left(\mathbb{S}^n \backslash T^{-1}(U_{\psi^c})\right) = 0,$$

where we used that $\mu(\mathbb{S}^n \backslash T^{-1}(U_{\psi^c})) = \nu(\mathbb{S}^n \backslash U_{\psi^c})$ since $\nu = T\#\mu$. Hence the set of points $x \in \mathbb{S}^n$ such that ψ is differentiable at x and ψ^c is differentiable at $T(x)$ is of full measure, i.e. $\mu(V) = 1$. Using Proposition 3.4.7 and the definition of T we see that for $x \in V$,

$$\begin{aligned} 0 &= c(x, T(x)) - \psi(x) - \psi^c(T(x)) \\ &= c(T(x), x) - \psi^c(T(x)) - \psi^{cc}(x), \end{aligned}$$

where $T(x)$ is uniquely determined by this equation. Using the symmetry of c, i.e. $c(x,y) = c(y,x)$, we also get that x is uniquely determined by $T(x)$, and therefore, by the definition of $T^\dagger$, $T^\dagger(T(x)) = x$ for $x \in V$. Replacing T by $T^\dagger$ and vice versa, we also get $T(T^\dagger(y)) = y$ μ-a.e. □

Corollary 3.4.12 (Polar factorization of spherical maps)**.** *Let $c(x,y) = \lambda(d(x,y))$ be a strictly convex* cost function. Let $S : \mathbb{S}^n \to \mathbb{S}^n$ be a Borel map and μ a probability measure on $\mathbb{S}^n$ absolutely continuous with respect to σ^n. Define $\nu := S\#\mu$, then, by Theorem 3.4.9, there exists a transport plan $T : \mathbb{S}^n \to \mathbb{S}^n$ defined by (3.20) for some c-concave map $\psi : \mathbb{S}^n \to \mathbb{R}$. If ν is also absolutely continuous with respect to σ^n, then there exists a measure preserving map $U : \mathbb{S}^n \to \mathbb{S}^n$, i.e. $U\#\mu = \mu$, such that $S = T \circ U$ μ-a.e. The maps T and U are μ-a.e. uniquely determined.*

Proof. Using Theorem 3.4.9 we get a c-concave ψ and a transport map $T : \mathbb{S}^n \to \mathbb{S}^n$. By Corollary 3.4.11 we get the inverse transport map $T^\dagger$ and define $U := T^\dagger \circ S$. Also there is $V \subseteq \mathbb{S}^n$ with $\nu(V) = 1$ such that for all $y \in V$ we have $T(T^\dagger(y)) = y$. Then

$$S(x) = (T \circ T^\dagger) \circ S(x) = (T \circ U)(x)$$

holds for all $x \in S^{-1}(V)$, i.e. μ-a.e. For $f \in \mathcal{C}(\mathbb{S}^n)$ we get

$$\int f \circ U d\mu = \int f \circ T^* \circ S \, d\mu = \int f \circ T^* d\nu = \int f d\mu$$

and therefore $U\#\mu = \mu$.
Thus we only need to show that T and U are μ-a.e. uniquely determined. Let T' be another optimal transport map such that $T'\#\mu = \nu$ and let U' be a measure preserving map such that $S = T' \circ U'$ holds on $W_1 \subseteq \mathbb{S}^n$ with $\mu(W_1) = 1$. By Theorem 3.4.9 there is $W_2 \subseteq \mathbb{S}^n$ with $\mu(W_2) = 1$ such that $T'(x) = T(x)$ for $x \in W_2$. Finally let $W_3 \subseteq \mathbb{S}^n$ be the set of all x such that $T^\dagger \circ T(x) = x$, then $\mu(W_3) = 1$ by Corollary 3.4.11. Put

$$B := W_1 \cap (U')^{-1}(W_2) \cap (U')^{-1}(W_3).$$

Then for $x \in B$, we have

$$S(x) \overset{x \in W_1}{=} T'(U'(x)) \overset{U'(x) \in W_2}{=} T(U'(x)). \tag{$\star$}$$

Thus

$$U(x) \overset{\text{def}}{=} T^\dagger(S(x)) \overset{(\star)}{=} T^\dagger(T(U'(x))) \overset{U'(x) \in W_3}{=} U'(x)$$

and since

$$\mu(\mathbb{S}^n \backslash B) \leq \underbrace{\mu(\mathbb{S}^n \backslash W_1)}_{=0} + \underbrace{\mu(\mathbb{S}^n \backslash (U')^{-1}(W_2))}_{\mu(\mathbb{S}^n \backslash W_2) = 0} + \underbrace{\mu(\mathbb{S}^n \backslash (U')^{-1}(W_3))}_{\mu(\mathbb{S}^n \backslash W_3) = 0} = 0$$

we have $U = U'$ μ-a.e. □

Remark 3.4.13. We should mention here the polar factorization Theorem for Euclidean maps due to Brenier [Bre91] and its extension to the setting of Riemannian manifolds due to McCann [McC01].

List of symbols

$\mathbb{Z}$	integers $\{\ldots,-2,-1,0,1,2\ldots\}$
$\mathbb{N}$	natural numbers $\{0,1,2\ldots\}$
$\mathbb{Q}$	rational numbers
$\mathbb{R}$	real numbers
$\overline{A}$	closure of a set A
dom	effective domain
$\mathrm{int}(A)$	interior of a set A
∂A	boundary of a set, $\partial A = \overline{A} \backslash \mathrm{int}(A)$
f^*	conjugate function of f
$\mathcal{C}$	continuous functions
$\partial^c f$	c-superdifferential of f
f^c	c-transform of f
$\mathcal{C}^n$	n-times continuously differentiable functions
epi	epigraph
$\mathrm{L}^1(X,\mu)$	μ-integrable functions on X
Lip	Lipschitz-continuous functions
cl	lower semi-continuous hull
$\delta^*(.\|C)$	support function of the convex set C
$\mathcal{B}$	σ-algebra of Borel sets
$\mathcal{P}$	Borel probability measures
$T\#\mu$	push-forward measure of μ by T
spt	support of a measure
W_p	Wasserstein distance of order p
$\mathbb{R}^n$	n-dimensional Euclidean space
∇f	gradient vector of a real-valued function f on $\mathbb{R}^n$
df	Jacobian-matrix of a function $f : \mathbb{R}^n \to \mathbb{R}^m$
$\mathcal{L}^n$	Lebesgue measure on $\mathbb{R}^n$

$\|.\|$	Euclidean vector norm in $\mathbb{R}^n$
$\langle .,. \rangle$	Euclidean vector product in $\mathbb{R}^n$
$\overline{\partial}_x f$	superdifferential of a real-valued function f in x
$\underline{\partial}_x f$	subdifferential of a real-valued function f in x
$\mathbb{S}^n$	n-dimensional sphere $\{x \in \mathbb{R}^{n+1} \mid \|x\| = 1\}$
σ^n	canonical volume measure on $\mathbb{S}^n$
$d(.,.)$	chord length metric on $\mathbb{S}^n$, $d(x,y) = \arccos(\langle x, y \rangle)$
$\mathrm{T}_x\mathbb{S}^n$	tangent vectors of $\mathbb{S}^n$ in x $\{y \in \mathbb{R}^{n+1} \mid \langle x, y \rangle = 0\}$
$\mathrm{U}_x\mathbb{S}^n$	normalized tangent vectors $\{y \in \mathrm{T}_x S^n \mid \|y\| = 1\}$

Bibliography

[BGMS09] Mathias Beiglböck, Martin Goldstern, Gabriel Maresch, and Walter Schachermayer, *Optimal and better transport plans*, J. Funct. Anal. **256** (2009), no. 6, 1907–1927. MR 2498564 (2010i:49075)

[Bre91] Yann Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Comm. Pure Appl. Math. **44** (1991), no. 4, 375–417. MR 1100809 (92d:46088)

[BS11] Mathias Beiglböck and Walter Schachermayer, *Duality for Borel measurable cost functions*, Trans. Amer. Math. Soc. **363** (2011), no. 8, 4203–4224. MR 2792985

[EG92] Lawrence C. Evans and Ronald F. Gariepy, *Measure theory and fine properties of functions*, Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1992. MR 1158660 (93f:28001)

[GM96] Wilfrid Gangbo and Robert J. McCann, *The geometry of optimal transportation*, Acta Math. **177** (1996), no. 2, 113–161. MR 1440931 (98e:49102)

[Kan42] Leonid V. Kantorovich, *On the translocation of masses.*, C. R. (Doklady) Acad. Sci. USSR **321** (1942), 199–201.

[Kan48] ———, *On a problem of monge (in russian).*, Uspekhi Mat. Nauk. **3** (1948), 225–226.

[McC01] Robert J. McCann, *Polar factorization of maps on Riemannian manifolds*, Geom. Funct. Anal. **11** (2001), no. 3, 589–608. MR 1844080 (2002g:58017)

[Roc70] R. Tyrrell Rockafellar, *Convex analysis*, Princeton Mathematical Series, No. 28, Princeton University Press, Princeton, N.J., 1970. MR 0274683 (43 #445)

[Sch93] Rolf Schneider, *Convex bodies: the Brunn-Minkowski theory*, Encyclopedia of Mathematics and its Applications, vol. 44, Cambridge University Press, Cambridge, 1993. MR 1216521 (94d:52007)

[Vil03] Cédric Villani, *Topics in optimal transportation*, Graduate Studies in Mathematics, vol. 58, American Mathematical Society, Providence, RI, 2003. MR 1964483 (2004e:90003)

[Vil09] ———, *Optimal transport*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 338, Springer-Verlag, Berlin, 2009, Old and new. MR 2459454 (2010f:49001)

Index