MONGE-KANTOROVICH AND TRANSPORTATION THEORY

COLIN YAO

Abstract. In this paper, we introduce the topic of transportation theory, focusing on the Monge-Kantorovich problem and its dual. The main result of the paper is a proof of the duality of the solutions to these two problems. Furthermore, we provide additional economic context to applications of this duality, and introduce a type of economic problem in which Monge-Kantorovich duality plays an unexpected, yet central role.

Contents

Main Paper 2
1. Introduction 2
1.1. Assumptions about the Reader 3
2. The Monge-Kantorovich Problem 3
2.1. The Monge Problem 3
2.2. Monge-Kantorovich 4
2.3. The Dual Problem 4
3. Weak Monge-Kantorovich 5
4. Strong Monge-Kantorovich 6
4.1. Part 1: Cyclically monotone transference plan in the discrete case 7
4.2. Part 2: Extension to cyclically monotone plan in general case 9
4.3. Part 3: Existence of lower-dimension dual 12
4.4. Part 4: Existence of Duality 15
4.5. Conclusion 19
5. Applications to Principal-Agent Problems 20
5.1. Overview 20
5.2. Incentive Design: Toy Example 20
5.3. Mathematical Footing 21
6. Acknowledgements 25
7. Bibliography 25
References 25

Appendix 25
Appendix A. Supplementary Theorems 25
A.1. Portmanteau Theorem 25
A.2. Law of Large Numbers for Empirical Measures 26
A.3. Prokhorov's Theorem 26
A.4. Fenchel Transform of c-concave Function 26
A.5. Lower and upper semi-continuity 27

Date: September 10, 2023.

A.6. Supremum of lower semi-continuous functions 27
A.7. Baire's Theorem 27
A.8. Integrability of Bounded and Measurable Functions 27
A.9. Convergence of Measurable Functions 27
Appendix B. Further Readings 27
B.1. Monopolist Profit Maximizing 28
B.2. Incentive Compatibility Theorems 28

. **Main Paper**

1. Introduction

Optimal transport is the study of how best to allocate different resources among different consumers of these resources. "Best" as a goal is vague, but generally taken to be the maximization of some kind of social welfare function. This is an important economics problem; the matching between producers and consumers which occurs every day is what allows for a market economy to efficiently allocate goods among all those that desire them.

As a field, optimal transport is relatively young. Gaspard Monge, an 18th-century French mathematician, first formulated the *Monge problem*, a forerunner of modern transportation problems which will be addressed in detail later. But roughly speaking, the Monge problem is to match producers to consumers one-to-one, such that the total cost of transportation between producers and consumers is minimized. Monge wrote at some length about this problem, but lacking modern mathematical tools, such as linear programming, he was unable to solve the problem.

Linear programming problems are maximization/minimization problems of specific form, and are the forerunners of more general optimal transport problems studied today. Essentially, every linear programming problem can be reduced to the problem of finding an n-dimensional vector $\mathbf{v}$ maximizing some objective function

$$c^T\mathbf{v}$$

(for c an n-dimensional vector of coefficients) subject to constraints of the form

$$A\mathbf{v} \leq b$$

for $k \times n$ matrix A and k-dimensional b. Once difficult and computationally intractable, the invention of the simplex method in 1947 by mathematician George Dantzig brought efficient solutions to this problem, and is still used today.

Also possibly of interest is the related *dual problem*: first conjectured by von Neumann when applying linear programming to zero-sum games, the dual problem for the above linear programming problem is the problem of finding k-dimensional $\mathbf{w}$ minimizing the function

$$b^T\mathbf{w}$$

given constraints

$$A^T\mathbf{w} \geq c.$$

Such a dual problem exists for every linear programming problem, and the solutions of the one coincide with the solutions of the other in a neat way which has cool economic implications. Duality for linear programs was established in 1948, and dual problems have been active subjects of research in the late 20th and early 21st centuries.

Thus, optimal transport in the discrete case was fairly well-studied during the later half of the 20th century, but conclusive study of the problem in its full generality did not occur until the late 1980s and onwards. The modern statement of the Monge-Kantorovich problem, which is the main subject of this paper, arises from this time period. Results from the Monge-Kantorovich problem and its dual solutions have had important effects on economics, physics, biology, and other fields as well.

1.1. **Assumptions about the Reader.** The results contained in this paper are intrinsically related to measure theory, and so the author will assume that the reader has at least some familiarity with the terminology of measure theory. But if one mentally substitutes the notions of "density" for measure, "everywhere that we care about" for "almost everywhere", and doesn't look too carefully at the underlying details, one can get by with a vague intuitional understanding.

Besides this, the reader should get accustomed to the idea of an analytic approach to probability, with probabilities defined by measures. Additional theorems are stated as needed in the text, with citations to complete proofs contained in the appendix. Otherwise, the rest of this paper should be comprehensible after a year of undergraduate analysis experience.

2. The Monge-Kantorovich Problem

First, we present Monge's original formulation of the problem, with updated terminology.

2.1. **The Monge Problem.** Let μ denote a probability measure on a space X, and let ν similarly denote a probability measure on X. Let $c : X \times X \to \mathbb{R}$ be a cost function. We denote a **transfer map** $\pi : X \to X$ such that π is one-to-one and for a set $S \subset X$, we have that $\mu(S) = \nu(\pi(S))$. Let $\Pi(\mu, \nu)$ denote the set of all transfer maps from μ to ν. Then our objective is to minimize

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} c(x, \pi(x)) d\mu.$$

2.1.1. *Commentary on the Monge Problem.* Monge was originally concerned with the problem of moving earth from heaps into holes. If we use μ, a measure, to represent the distribution of earth in above-ground heaps over the space X, and ν, another measure, to represent the distribution of vacancies underground in X, and we assume that there is exactly enough earth to flatten every heap and fill every hole, then we can normalize both μ and ν and use the terminology of probability measures. Our transfer map π then declares how every infinitesimal vertical cross-section of dirt is to be transferred from place to place, and c represents the cost involved with moving earth from place to place. Certain locations may be convenient to transfer earth to, or may be further away, or any combination of such

factors.

Nevertheless, while the physical interpretation of such a plan can be seen, actually finding such a plan remains a difficult problem to this day. The cases where $c(x,y) = |x-y|$ and $c(x,y) = (x-y)^2$ are fairly well studied and can be solved, but more exotic cost functions are less tractable.

Furthermore, the Monge problem may be ill-posed for certain distributions μ, ν. For example, for $X = \mathbb{R}$, $\mu = \delta_0$, $\nu = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2$ (where δ_k is the Dirac delta function centered at k), there are no admissible transport maps at all, because we cannot split the mass located at 0 to send it to 1 or 2, as such a map would not be one-to-one. Leonid Kantorovich would later recognize that relaxing this one-to-one condition would make the problem significantly more tractable and allow for a linear programming approach to be used. We now present the modern formulation of the Monge-Kantorovich problem.

2.2. **Monge-Kantorovich.** Let μ denote a probability measure on a space X, and let ν denote a probability measure on a space Y. Let $c : X \times Y \to \mathbb{R}$ be a cost function. We say that $\pi : X \times Y \to \mathbb{R}$ is a *transference plan* if it is a probability measure on $X \times Y$, and for arbitrary sets $S \subset X, T \subset Y$, we have that $\pi(S \times Y) = \mu(S)$ and $\pi(X \times T) = \nu(T)$. We let $\Pi(\mu, \nu)$ denote the set of all admissible transference plans. Then the problem is to minimize:

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x,y) d\pi.$$

2.2.1. *Commentary on Monge-Kantorovich.* The central idea of the relaxation is the probability distribution, which changed from Monge's idea of a one-to-one function which described something like a matching, into a probability distribution over the combined spaces. This allows for us to perform tricks like shifting half the mass of a point (x_0, y_0) to a different point (x_0, y_1), and shifting the same amount of mass from (x_1, y_1) to (x_1, y_0), thus preserving the amount of mass at points x_0, x_1, y_0, y_1 while materially changing the distribution. This ease of transfer is what makes this problem both more soluble and more general than Monge's original formulation.

If you close your eyes and squint, the Monge-Kantorovich problem as presented above looks a bit like a linear programming problem, with the constraint being that $\pi(X \times Y) = 1$. Indeed, if π assigns positive measure only to discrete points, this is in fact a linear programming problem. Thus, a natural question is to ask about the dual problem and its interpretation. Here it is:

2.3. **The Dual Problem.** Let $\psi : X \to \mathbb{R}$ and $\phi : Y \to \mathbb{R}$ be integrable functions such that for almost every $(x,y) \in X \times Y$ (outside a set of measure zero), $\psi(x) + \phi(y) \leq c(x,y)$. The dual problem is then the maximization:

$$\max_{\psi,\phi} \int_X \psi(x) d\mu + \int_Y \phi(y) d\nu.$$

2.3.1. *Commentary on the Dual Problem.* As this is a dual problem, the two solutions in fact should coincide:

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x,y) d\pi = \sup_{\psi,\phi} \left(\int_X \psi(x) d\mu + \int_Y \phi(y) d\nu \right). \tag{2.1}$$

This is known as Monge-Kantorovich duality, and demonstrating that this holds is the main content of this paper. Furthermore, we will include some examples of how it can be applied to solve certain problems outside of transportation theory.

Here is an economic interpretation of this statement: Suppose that X represents a space of bakeries (with bakeries positioned according to their characteristics, such as clientele and location), and Y represents a space of cafes serving baked goods. Bakeries produce bread according to the probability distribution μ (so for a finite number of bakeries, this is a probability distribution on point masses), and cafes demand bread according to the probability distribution ν. Acting on their own, the bakeries and cafes must incur some cost $c(x, y)$ to transport a unit of bread from bakery x to cafe y, and their problem is to find the distribution π such that the total cost of transporting bread, $\int_{X\times Y} c(x, y)d\pi$, is minimized.

Now, suppose a transportation company offers to take care of the transportation between bakeries and cafes. They charge a flat fee for pickup at a bakery: $\psi(x)$ per unit, depending on the bakery x, and another fee for delivery to a cafe: $\phi(y)$ per unit, depending on the target cafe y. The company guarantees that their prices are competitive: $\psi(x) + \phi(y) \leq c(x, y)$ for almost every pair (x, y), so it is always worth the bakery/cafe's while to use the transportation company's services.

Then, the duality statement implies that for the transportation company, there is some pair of pricings ψ, ϕ such that the transportation company earns essentially as much as the bakeries/cafes were spending on their own. (So essentially all the slack in the market is taken up by the transportation company.) This is a non-obvious statement, but it has the benefit of being true. We will proceed by first proving a weaker statement, and then proving the result in full detail. The proof of the weak statement is adapted from Alfred Galichon's *Optimal Transport Methods in Economics* [13].

3. Weak Monge-Kantorovich

Take the assumptions as in Section 2.2. Instead of demonstrating equality as in Equation (2.1):

$$\inf_{\pi\in\Pi(\mu,\nu)} \int_{X\times Y} c(x, y)d\pi = \sup_{\psi,\phi} \left(\int_X \psi(x)d\mu + \int_Y \phi(y)d\nu \right)$$

we show a weaker inequality first:

$$\inf_{\pi\in\Pi(\mu,\nu)} \int_{X\times Y} c(x, y)d\pi \geq \sup_{\psi,\phi} \left(\int_X \psi(x)d\mu + \int_Y \phi(y)d\nu \right). \tag{3.1}$$

Informally, this claims that the inequality

$$\psi(x) + \phi(y) \leq c(x, y)$$

which holds pointwise at every (x, y) can be extended to the integral over the entire space $X \times Y$.

Proof. As π satisfies $\pi(S \times Y) = \mu(S)$ for subsets S of X, we claim that

$$\int_X \psi(x)d\mu = \int_{X\times Y} \psi(x)d\pi,$$

as the measures assigned are the same. Similarly,

$$\int_Y \phi(y)d\nu = \int_{X\times Y} \phi(y)d\pi.$$

Then, as we know that $c(x,y) \geq \psi(x) + \phi(y)$ almost everywhere, we know that the following inequality holds:

$$\int_{X\times Y} c(x,y)d\pi \geq \int_{X\times Y} \psi(x) + \phi(y)d\pi.$$

Thus, combining these statements, we have that

$$\int_{X\times Y} c(x,y)d\pi \geq \int_X \psi(x)d\mu + \int_Y \phi(y)d\nu,$$

Taking the infimum on the left-hand side and the supremum on the right-hand side, we attain Equation (3.1), as desired. □

4. Strong Monge-Kantorovich

In its strong form, Monge-Kantorovich duality claims that the two quantities concerned are equal, namely that

$$\inf_{\pi\in\Pi(\mu,\nu)} \int_{X\times Y} c(x,y)d\pi = \sup_{\psi,\phi} \left(\int_X \psi(x)d\mu + \int_Y \phi(y)d\nu \right),$$

as in Equation (2.1). While the duality can be applied to any lower semi-continuous cost functions, we will restrict the cases under consideration to continuous nonnegative cost functions. The proof of the theorem will be laid out in four parts as follows:

(1) Demonstrating that for discrete measures μ and ν, a coupling which has μ and ν as marginal probabilities and is *minimal* in some sense does exist.
(2) Using marginals with discrete measure to extend via a limit to a pairing in arbitrary measures, creating a coupling which is minimal in the same sense.
(3) Showing the existence of one of the one-sided cost functions ψ, ϕ with some desirable properties.
(4) Using the other corresponding one-sided cost function to show that duality does in fact hold.

This approach is adapted from the proof presented by Cedric Villani in his *Optimal Transport: Old and New* [14].

Before we begin, a full and precise statement of the Monge-Kantorovich theorem:

Theorem 4.1. *Let $(X,\mu),(V,\nu)$ be Polish probability spaces. Let $c : X \times Y \to \mathbb{R}^+$ be a continuous cost function. Then, the following equality holds:*

$$\min_{\pi\in\Pi(\mu,\nu)} \int_{X\times Y} c(x,y)d\pi(x,y) = \sup_{\psi,\phi} \left(\int_X \psi(x)d\mu(x) + \int_Y \phi(y)d\nu(y) \right)$$

over all $\psi, \phi \in C_b(X), C_b(Y)$ and for all x, y, $\psi(x) + \phi(y) \leq c(x,y)$.

(Polish spaces are separable, completely metrizable spaces, but for most purposes, a compact subset of $\mathbb{R}^n$ with the Lebesgue measure will suffice. As described in Section 2.2, $\Pi(\mu,\nu)$ is the set of admissible transference plans, namely those with marginals μ, ν on X, Y. The set $C_b(X)$ is the set of continuous bounded functions on X, and the same for $C_b(Y)$.)

4.1. **Part 1: Cyclically monotone transference plan in the discrete case.** Before we begin, we establish some definitions:

4.1.1. *Preliminaries.*

Definition 4.2. *Let X, Y be arbitrary sets and $c : X \times Y \to \mathbb{R}$ a function. A subset $\Gamma \subset X \times Y$ is defined to be c-**cyclically monotone** if, for any $N \in \mathbb{N}$, any collection of points $(x_1, y_1), \cdots, (x_N, y_N) \in \Gamma$, the following relation holds:*

$$\sum_{i=1}^{N} c(x_i, y_i) \leq \sum_{i=1}^{N} c(x_i, y_{i+1}) \tag{4.3}$$

*(with the convention that $y_{N+1} = y_1$). A transference plan (a coupling between X and Y, which we model by a joint probability measure on the space $X \times Y$) is said to be c-**cyclically monotone** if it is concentrated (has support on, excepting a set of measure zero) on a c-cyclically monotone set.*

This idea of cyclical monotonicity provides the foundation for comparison between the cost-minimizing and profit-maximizing cases. Intuitively speaking, if a transference plan π is not c-cyclically monotone, then as there exists some finite cycle on which π assigns positive mass to every point, we can arrange for a change in π to a new transference plan $\tilde{\pi}$ by decreasing the weight of each (x_i, y_i) and increasing the weight of each (x_i, y_{i+1}) by the same amount. The new joint probability distribution $\tilde{\pi}$ still has the same marginals μ, ν as the original π, so the infimum side of the duality is unchanged, but the supremum side of the duality decreases. So if we want to compare these two, we should make sure that the supremum side is in no position to change.

Note that cyclic monotonicity does not necessarily imply general optimality, although an optimal plan must be cyclically monotone (as it would otherwise be able to be improved). But a cyclically monotone plan at least cannot be easily improved.

4.1.2. *Statement and Proof.* Before we attempt to find cyclically monotone transference plans for general measures, we first solve the problem for the specific case of measures with a finite number of discrete point masses. From this solution, we aim to approximate the continuous case with a sequence of discrete cases. But we should hope that even in the discrete case, we can find a cyclically monotone transference plan. Thus, we wish to show the following proposition:

Proposition 4.4. *Let $\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ and $\nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}$, where the cost $c(x_i, y_i)$ is finite at every point (x_i, y_i). Then there exists a c-cyclically monotone transference plan.*

Proof. Consider the space of all transference plans. There are only $n \times n$ points in $X \times Y$ which can be assigned positive mass, so we can associate the transference plan with some $n \times n$ matrix A, with each element a_{ij} referring to the mass at the point (x_i, y_j). (Note that such a matrix still must obey the requirements for the marginal distribution, so the sum of each column and the sum of each row must total $\frac{1}{n}$.)

Now, the corresponding Monge-Kantorovich problem is the following:

$$\inf \sum_{i,j} a_{ij} \cdot c(x_i, y_j)$$

where the infimum is over all $n \times n$ matrices (with the right column / row sums). But if we instead look at the space of $n \times n$ matrices as $\mathbb{R}^{n^2}$, each of the row/column sum constraints defines a hyperplane. The equation

$$\sum_{j=1}^{n} a_{0j} = \frac{1}{n},$$

which constrains the first row to have sum $\frac{1}{n}$, defines a hyperplane of degree $n-1$. The further constraint (due to measure) that each entry is nonnegative is equivalent to a restraint along another hyperplane for each of the n dimensions involved. For example, for $n = 3$, the constraint above produces a plane which intersects the points $(\frac{1}{3}, 0, 0), (0, \frac{1}{3}, 0), (0, 0, \frac{1}{3})$. The non-negativity constraint reduces this plane to the 2D triangle embedded in $\mathbb{R}^3$ with these points as vertices.

In this way, each of the constraints restricts the total space of matrices to the intersection of these convex polytopes, ultimately producing a convex polytope. This polytope is nonempty, as $\frac{1}{n} \cdot I$ is located within the polytope, where I is the $n \times n$ identity matrix, so the space of permissible matrices is compact. Then, we are minimizing this linear function over a compact set, so the infimum is attained by some matrix A. We can then define the transference plan π as

$$\pi := \frac{1}{n} \sum_{i,j} a_{ij} \cdot \delta_{x_i, y_j}.$$

This π is c-cyclically monotone: if it weren't, then there would exist some subsequences $\{i_k\}$ and $\{j_k\}$ such that

$$\sum_{l=1}^{k} c(x_{i_l}, y_{j_l}) > \sum_{l=1}^{k} c(x_{i_l}, y_{j_{l+1}}).$$

(We have the convention that $j_{k+1} = j_1$, similarly to above.) Furthermore, at each point (x_{i_l}, y_{j_l}), the constant $a_{i_l j_l}$ must be strictly positive, as π must assign positive mass to this point for it to violate c-cyclic monotonicity. Then, letting $b = \min_{l \in [k]} a_{i_l j_l}$, we have that $b > 0$ and for each l, $a_{i_l j_l} - b \geq 0$.
We can then define a new transference plan $\tilde{\pi}$ as follows:

$$\tilde{\pi} := \pi - \frac{1}{n} \sum_{l=1}^{k} b \cdot \delta_{x_{i_l}, y_{j_l}} + \frac{1}{n} \sum_{l=1}^{k} b \cdot \delta_{x_{i_l}, y_{j_{l+1}}}.$$

By what we just showed above, this is in fact a valid transference plan: the i_lth row is decreased by $\frac{b}{n}$ and increased by $\frac{b}{n}$, so the row/column sums are still $\frac{b}{n}$ as before. Each pair (x_i, y_j) still has positive mass associated with it as well. But by our original assumption, π minimizes the cost function

$$C(\pi) = \sum_{i,j} a_{ij} \cdot c(x_i, y_j).$$

Plugging $\tilde{\pi}$ into the cost function, we have

$$C(\tilde{\pi}) = \sum_{i,j} a_{ij} \cdot c(x_i, y_j) - \frac{1}{n} \sum_{l=1}^{k} b \cdot c(x_{i_l}, y_{j_l}) + \frac{1}{n} \sum_{l=1}^{k} b \cdot c(x_{i_l}, y_{j_{l+1}}).$$

But as

$$\sum_{l=1}^{k} c(x_{i_l}, y_{j_l}) > \sum_{l=1}^{k} c(x_{i_l}, y_{j_{l+1}}) \tag{4.5}$$

from the assumption that π was not c-cyclically monotone, Equation (4.5) rearranges to

$$-\frac{1}{n} \sum_{l=1}^{k} b \cdot c(x_{i_l}, y_{j_l}) + \frac{1}{n} \sum_{l=1}^{k} b \cdot c(x_{i_l}, y_{j_{l+1}}) < 0$$

and so

$$C(\tilde{\pi}) < C(\pi).$$

We assumed that π minimized the cost function, so this is a contradiction. Thus, π must be c-cyclically monotone.

□

4.2. **Part 2: Extension to cyclically monotone plan in general case.** As the Monge-Kantorovich problem in its full generality deals with probability measures μ, ν in a Polish probability space, we would like to use the result from the discrete case, where we had only point measures, and extend it to more general measures. To do so, we first introduce some pre-existing results from probability theory, which we will not immediately provide proofs for. (Proofs or citations to complete proofs will be included in the appendix.)

4.2.1. *Preliminaries.*

Definition 4.6. *Let (S, Σ) be a space and its σ-algebra, and let $\{P_n\}$ be a sequence of bounded positive probability measures on (S, Σ). The sequence $\{P_n\}$* ***converges weakly*** *to a probability measure P if:*

$$\lim_{n\to\infty} \int_S |f| dP_n = \int_S |f| dP$$

for all bounded continuous functions f.

Theorem 4.7. *(Portmanteau Theorem) A sequence of measures $\{\mu_n\}$ converges weakly to μ if and only if*

$$\limsup_{n\to\infty} \mu_n(T) \leq \mu(T)$$

for all closed sets T.

Proof in Appendix A.1.

Theorem 4.8. *(Law of Large Numbers for Empirical Measures) Let $\{X_n\}$ be a sequence of i.i.d. (independent, identically distributed) random variables on the space X with distribution according to some probability measure μ. Then the sample probability measure for n points, μ_n, converges to the true probability measure μ:*

$$\mu_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \to \mu$$

where the convergence is weak convergence as in Definition 4.6.

Proof in Appendix A.2.

Definition 4.9. *We call a measure m on a space S **tight** if for any $\epsilon > 0$, there exists some compact set K_ϵ of S such that $m(S \setminus K_\epsilon) < \epsilon$. A collection of measures M is **tight** if for any $\epsilon > 0$, there exists compact K_ϵ such that $m_\alpha(S \setminus K_\epsilon) < \epsilon$ for every measure m_α in M.*

Definition 4.10. *A space S is **sequentially compact** if every sequence in S has a convergent subsequence which converges to a point in S.*

Theorem 4.11. *(Prokhorov's Theorem) Let S be a separable metric space, and let $\mathcal{P}(S)$ be the set of all probability measures defined on S. Then a collection $M \subset \mathcal{P}(S)$ of probability measures is tight if and only if the closure of M is sequentially compact in the space $\mathcal{P}(S)$ with respect to weak convergence.*

Proof in Appendix A.3.

Definition 4.12. *Let X, Y be two spaces. If μ is a Borel measure on X, and $T : X \to Y$ is a Borel function, we define the **pushforward** of μ by T as*

$$(T_{\#}\mu)(A) := \mu(T^{-1}(A))$$

for all measurable sets A in Y. Note that $T_{\#}\mu$ is a Borel measure on Y.

Definition 4.13. *Let π be a probability distribution with respect to two variables $x \in X$ and $y \in Y$. Its **marginal** on X is the measure $f_{\#}\pi$, where $f(x, y) = x$. We define the **marginal** on Y similarly: $g_{\#}\pi$, where $g(x, y) = y$.*

Definition 4.14. *Let μ be a measure on X and ν be a measure on Y. We define the **product measure** $\mu \otimes \nu$ as the measure on $X \times Y$ such that for any measurable sets B_1, B_2 in X, Y respectively,*

$$(\mu \otimes \nu)(B_1 \times B_2) = \mu(B_1) \cdot \nu(B_2).$$

In the case where X and Y are the same space and μ and ν are the same measure, we write $\mu \otimes \mu$ as $\mu^{\otimes 2}$, and generally

$$\underbrace{\mu \otimes \mu \otimes \cdots \otimes \mu}_{n \text{ times}} = \mu^{\otimes n}.$$

Definition 4.15. *We say that measure μ on a space X is **concentrated** on some set $S \subset X$ if $\mu(X \setminus S) = 0$.*

4.2.2. *Lemma (tightness of transference).* Now, before we begin the proof, we introduce a technical lemma specific to this problem.

Lemma 4.16. *Let X, Y be Polish spaces, $M \subset \mathcal{P}(X)$, $N \subset \mathcal{P}(Y)$, (where $\mathcal{P}(X)$ is the set of probability measures on X as before) and let M, N be tight. Let $\Pi(M, N)$ be the set of all transfer plans π such that the marginals of π with respect to X, Y lie in M, N respectively. Then this set $\Pi(M, N)$ is tight in $\mathcal{P}(X \times Y)$.*

Proof. Choose any $\epsilon > 0$. Then, as M, N are tight, there exist compact K_ϵ, L_ϵ such that for any $\mu \in M$, $\mu(X \setminus K_\epsilon) < \frac{\epsilon}{2}$ and for any $\nu \in N$, $\nu(Y \setminus L_\epsilon) < \frac{\epsilon}{2}$.

Now, take some $\pi \in \Pi(M, N)$, with marginals on X, Y of μ, ν respectively. ($\mu \in M, \nu \in N$.) The set $(X \times Y) \setminus (K_\epsilon \times L_\epsilon)$ can be decomposed into two (not necessarily disjoint) sets $(X \setminus K_\epsilon) \times Y$ and $X \times (Y \setminus L_\epsilon)$, so

$$\pi(X \times Y) \setminus (K_\epsilon \times L_\epsilon) \leq \pi((X \setminus K_\epsilon) \times Y) + \pi(X \times (Y \setminus L_\epsilon)).$$

Considering $\pi((X \setminus K_\epsilon) \times Y)$, let $f(x, y) = x$ as in Definition 4.13. Then as

$$(X \setminus K_\epsilon) = f^{-1}((X \setminus K_\epsilon) \times Y)$$

we have that

$$\begin{aligned}\pi(f^{-1}(X \setminus K_\epsilon)) &= f_{\#}\pi(X \setminus K_\epsilon) \\ &= \mu(X \setminus K_\epsilon) \\ &\leq \frac{\epsilon}{2}.\end{aligned}$$

Similarly, we have that $\pi(X \times (Y \setminus L_\epsilon)) \leq \frac{\epsilon}{2}$ as well, so $\pi((X \times Y) \setminus (K_\epsilon \times L_\epsilon)) \leq \epsilon$. The Cartesian product of compact sets $K_\epsilon \times L_\epsilon$ is compact as well, so we have found a compact set satisfying the requirements for tightness, and so π is tight. This proof works independent of which π is chosen, with the same K_ϵ, L_ϵ, so the entire set $\Pi(M, N)$ is tight, as desired. □

This lemma will be combined with Prokhorov's Theorem (Theorem 4.11) to demonstrate convergence of a sequence of probability measures to a single probability measure later on. With this lemma in hand, we can move forward with the proof of this section.

4.2.3. *Statement and Proof.* We wish to show the following proposition:

Proposition 4.17. *For continuous cost function c, there exists a cyclically monotone transference plan π with marginals on X and Y of μ, ν respectively.*

Proof. Consider two sequences of i.i.d. random variables $\{x_n\}$ and $\{y_n\}$ with distributions according to probability distributions μ, ν on X, Y respectively. Define the sample probability measure μ_n for the first n elements of $\{x_n\}$ by

$$\mu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$

as in Theorem 4.8 above, and ν_n for $\{y_n\}$ similarly. Then, by Theorem 4.8, we know that μ_n converges to μ weakly, and ν_n converges to ν weakly as well. Then, as the sequence $\{\mu_n\}$ converges to the probability measure μ, it is sequentially compact, as every subsequence must also converge to μ. Then, by Prokhorov's Theorem (Theorem 4.11), the sequence $\{\mu_n\}$ is tight, and similarly $\{\nu_n\}$ is tight as well.

For each n, let π_n be the cyclically monotone transference plan between μ_n and ν_n, which we showed to exist in Section 4.1. By Lemma 4.16, the sequence $\{\pi_n\}$ is tight as well, so by Prokhorov again (Theorem 4.11) this sequence $\{\pi_n\}$ must be sequentially compact, and thus has a subsequence converging to some π. If we let the subsequence be $\{\kappa_n\}$, we have that

$$\int_{X\times Y} h d\kappa_n \to \int_{X\times Y} h d\pi$$

for every bounded continuous $h(x, y)$. If we plug in some function $f(x)$ depending only on x, we have that

$$\int_{X\times Y} f d\kappa_n = \int_X f d\mu_n,$$

so

$$\lim_{n\to\infty}\int_{X\times Y} f d\kappa_n = \lim_{n\to\infty}\int_X f d\mu_n = \int_X f d\mu.$$

So as this holds for any $f(x)$, and similarly for any $g(y)$, π has marginals on X, Y of μ, ν respectively, as desired.

We now show that π defined as above is cyclically monotone:

Let $C(N)$ be the set of N-cycles in $(X \times Y)^N$ such that for a point $p = ((x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N))$ in $C(N)$, we have that

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

as in Equation (4.3) (with the same convention $y_{N+1} = y_1$). We know that every κ_n is cyclically monotone, so for any point $q = ((x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N))$ **not** in $C(N)$, κ_n cannot assign positive mass to each point $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$, or else κ_n is not cyclically monotone. Thus, for every κ_n, we can say that the mass of $\kappa_n^{\otimes N}$ is concentrated on $C(N)$.

As the cost function c is continuous, $C(N)$ is closed, as every convergent sequence of cyclically monotone N-cycles must converge to a cyclically monotone N-cycle. But then by the Portmanteau Theorem (Theorem 4.7), we must have that

$$\limsup_{n\to\infty} \mu_n(C(N)) \leq \mu(C(N)),$$

and so as $\mu_n(C_N) = 1$ for all μ_n, $\mu(C(N)) = 1$ as well. Thus μ is concentrated only on $C(N)$ for each N, and so μ must be cyclically monotone. □

4.3. **Part 3: Existence of lower-dimension dual.** Similar to before, we must introduce some technical definitions in order to properly define the desired statement in this part.

4.3.1. *Preliminaries.*

Definition 4.18. *Let X, Y be spaces, and $c : X \times Y \to \mathbb{R}$. A function $\psi : X \to \mathbb{R}$ is c-**concave** if there exists some $\phi : Y \to \mathbb{R}$ such that for all x,*

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

This definition is somewhat abstract, so here's the intuition: In the case where $X, Y = \mathbb{R}$, $c(x, y)$ is a function from $\mathbb{R}^2$ to $\mathbb{R}$. (You can visualize this as an irregular surface of varying height over some area, like $[0, 1] \times [0, 1]$.) The function ψ defines a surface with a height which varies only with respect to x, and not to y. On a line parallel to the y-axis, ϕ has the same value at every point. (Think of this as a piece of corrugated cardboard, with the corrugations all running in the same direction.) If ψ is c-concave, then after overlaying some $\phi(y)$ with corrugations in the orthogonal direction (parallel to the x-axis), the resulting surface $(\psi + \phi)$ is tangent to that defined by c at some point (x, y) for every x.

Equipped with this definition of c-concavity, we now examine a related concept:

Definition 4.19. *The c-conjugate of ψ, written $\psi^c(y)$, is defined:*

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

(This is also known as the Legendre-Fenchel transform, among other names.) Note that ψ^c is precisely that ϕ for which we can show that ψ is c-concave.

Definition 4.20. *The c-subdifferential of ψ, denoted $\partial_c \psi$, is the set of points $(x,y) \in X \times Y$ satisfying*

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

The c-conjugate satisfies some nice properties, like how for ψ c-convex, $(\psi^c)^c = \psi$. Just as ψ is a function on X, ψ^c is intended to be its counterpart on Y, as a function that "fills gaps" where ψ provides opportunities between itself and c.

These points (x,y) where gaps can be completely filled are grouped into the c-subdifferential, where $\psi(x) + \phi(y) = c(x,y)$ holds. Note that at points (x,y) in $\partial_c \psi$, the infimum in the definition of ψ^c is actually achieved by the x-coordinate of (x,y).

Definition 4.21. *The **support** of a measure or a function m is the smallest closed set such that the set of all points outside the support where m is nonzero has measure zero.*

4.3.2. *Statement.* Now, we present the main statement that we wish to show in this part:

Proposition 4.22. *For c continuous, π c-cyclically monotone, there exists a c-convex ψ such that Support$(\pi) \subset \partial_c \psi$.*

In the previous part, we showed the existence of a cyclically monotone π. Now, based on this π, we find a ψ such that ψ "fits tightly" under c in the way of c-convexity. Furthermore, we want all the points where the optimal matching π assigns positive mass to be contained in the c-subdifferential of ψ: so at every point that we care about (has positive mass in the optimal matching) the sum of $\psi(x)$ and $\psi^c(y)$ is equal to $c(x,y)$. Now we see signs of our ultimate goal: ψ, ψ^c are the functions in the dual problem, and having them be close to/equal to c on π would be really nice, as then equality of the integrals in Equation (2.1) would follow.

Proof. Let $\Gamma = \text{Support}(\pi)$. Pick any $(x_0, y_0) \in \Gamma$. Now, we define ψ:

(4.23)
$$\psi(x) := \inf_{m \in \mathbb{N}} \inf((c(x_1,y_0) - c(x_0,y_0)) + (c(x_2,y_1) - c(x_1,y_1)) + \cdots + (c(x,y_m) - c(x_m,y_m)) : (x_1,y_1), \cdots, (x_m,y_m) \in \Gamma).$$

Informally, $\psi(x)$ represents the maximum possible difference that can be created by taking a finite cycle of cost differences, and replacing the quantity $c(x_0,y_m)$ with $c(x,y_m)$ instead.

For the length-2 cycle $(x_1,y_1) = (x_0,y_0)$, the difference is zero, so we know that $\psi(x_0) \leq 0$. But as each point in any cycle must be in Γ, and we showed in the previous part (Section 4.2) that Γ is cyclically monotone, all such cycles must satisfy

$$\sum_{i=0}^{m} c(x_{i+1}, y_i) - c(x_i, y_i) \geq 0.$$

Thus, $\psi(x_0) = 0$.

Now, we consider the consequences of allowing y_m to be a choice variable as well. We can rewrite $\psi(x)$ as

(4.24)
$$\psi(x) = \inf_{y\in Y} \inf_{m\in\mathbb{N}} \inf((c(x_1,y_0)-c(x_0,y_0))+(c(x_2,y_1)-c(x_1,y_1))+\cdots+(c(x,y_m)-c(x_m,y)) : (x_1,y_1),\cdots,(x_m,y)\in\Gamma).$$

(Note the change from the definition of ψ in Equation (4.23) above: we have $(x_m, y) \in \Gamma$ here instead of $(x_m, y_m) \in \Gamma$.)

This is getting unwieldy. Let's take the expression in the last two infima of Equation (4.24) and rewrite it as a function $\zeta(y)$ of y:

$$\zeta(y) := \inf_{m\in\mathbb{N}} \inf((c(x_1,y_0)-c(x_0,y_0))+(c(x_2,y_1)-c(x_1,y_1))+\cdots+(c(x,y_m)-c(x_m,y_m)) : (x_1,y_1),\cdots,(x_m,y)\in\Gamma).$$

Then we can write $\psi(x)$ as the much more concise

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

But then notice that this is precisely the definition of c-concavity in Definition 4.18, so ψ is in fact c-concave.

This last manipulation isn't just sleight of hand: ψ was especially constructed to be able to be manipulated into this form. (The term $c(x, y_m)$ in eq. (4.23) lets y_m float, and we get to choose the (x_m, y_m) minimizing.)

Now, pick any $(\tilde{x}, \tilde{y}) \in \Gamma$. If we fixed the (x_m, y_m) in the original definition of ψ as $(\tilde{x}, \tilde{y})$, we would have:

$$\psi(x) \le \inf_{m\in\mathbb{N}} \inf((c(x_1,y_0)-c(x_0,y_0))+(c(x_2,y_1)-c(x_1,y_1))+\cdots+(c(\tilde{x},y_{m-1})-c(x_{m-1},y_{m-1}))+(c(x,\tilde{y})-c(\tilde{x},\tilde{y}) (x_1,y_1),\cdots,(x_{m-1},y_{m-1})\in\Gamma).$$

(We've essentially restricted the set of possible cycles to those which have $(\tilde{x}, \tilde{y})$ as their last point before repeating again at (x_0, y_0).) Let's now consider what $\psi(\tilde{x})$ is:

$$\psi(\tilde{x}) = \inf_{m\in\mathbb{N}} \inf((c(x_1,y_0)-c(x_0,y_0))+(c(x_2,y_1)-c(x_1,y_1))+\cdots+(c(\tilde{x},y_m)-c(x_m,y_m))) : (x_1,y_1),\cdots,(x_m,y_m)\in\Gamma).$$

We're taking the infimum over all $m \in \mathbb{N}$, so we can reindex:

$$\psi(\tilde{x}) = \inf_{m\in\mathbb{N}} \inf((c(x_1,y_0)-c(x_0,y_0))+(c(x_2,y_1)-c(x_1,y_1))+\cdots+(c(\tilde{x},y_{m-1})-c(x_{m-1},y_{m-1}))) : (x_1,y_1),\cdots,(x_{m-1},y_{m-1})\in\Gamma).$$

But then if we're taking the infimum over n points in Equation (4.23), we can take the supremum over $n-1$ points and then another point, so the above expression

for $\psi(x)$ can be rewritten:

$$\psi(x) = \inf_{(\tilde{x},\tilde{y})\in\Gamma} \inf_{m\in\mathbb{N}} \inf((c(x_1,y_0)-c(x_0,y_0))+(c(x_2,y_1)-c(x_1,y_1))+\cdots+(c(x,\tilde{y})-c(\tilde{x},\tilde{y}))) : \\ (x_1,y_1),\cdots,(x_{m-1},y_{m-1})\in\Gamma)$$

And this we can condense:

$$\psi(x) = \inf_{(\tilde{x},\tilde{y})\in\Gamma} \psi(\tilde{x}) + c(x,\tilde{y}) - c(\tilde{x},\tilde{y})$$

So for any $(\tilde{x},\tilde{y})\in\Gamma$, we have that

$$\psi(x) \le \psi(\tilde{x}) + c(x,\tilde{y}) - c(\tilde{x},\tilde{y}).$$

Rearranging to

$$-\psi(x) + c(x,\tilde{y}) \ge -\psi(\tilde{x}) + c(\tilde{x},\tilde{y})$$

we can take the infimum of the left-hand side over x to yield

$$\inf_{x\in X}(-\psi(x) + c(x,\tilde{y})) \ge -\psi(\tilde{x}) + c(\tilde{x},\tilde{y}).$$

But recall now from the definition of the c-conjugate in Definition 4.19 that the left-hand side is actually $\psi^c(y)$. Thus, this finally rearranges to $\psi^c(\tilde{y}) + \psi(\tilde{x}) \ge c(\tilde{x},\tilde{y})$. From the definition of the c-conjugate,

$$\psi^c(\tilde{y}) \le -\psi(\tilde{x}) + c(\tilde{x},\tilde{y}),$$

so combining these two statements, we have that equality must in fact hold, and so $\psi^c(\tilde{y}) + \psi(\tilde{x}) = c(\tilde{x},\tilde{y})$ for all points $(\tilde{x},\tilde{y})$ in Γ. Thus, we have shown that

$$\Gamma \subset \partial_c\psi,$$

as desired. □

As the reader may suspect, this ψ so defined is in fact the ψ from the duality statement of the Monge-Kantorovich problem above. Additionally, ψ^c is ϕ. Demonstrating that this is true is the objective of the next section.

4.4. **Part 4: Existence of Duality.** In the last part, we introduced ψ, and then proceeded to show that it satisfied the desirable property

$$\Gamma \subset \partial_c\psi,$$

as well as being c-concave. Here, we deal with several other issues, such as showing that ψ and ψ^c are measurable and integrable, so that we can take the integral and be confident that it doesn't blow up in an unexpected way.

4.4.1. *Preliminaries.* There is another property of c-convexity which makes it quite valuable for our purposes.

Theorem 4.25. *If function ψ is c-concave, then $(\psi^c)^c = \psi$. (ψ^c is the c-transform of ψ, as in Definition 4.19.)*

Proof in Appendix A.4.

Definition 4.26. *A function $f : X \to \mathbb{R}$ is **lower semi-continuous** at $x_0 \in X$ if for every real $y < f(x_0)$, there exists a neighborhood U of x_0 such that $f(x) > y$ for all $x \in U$.*

Recall from the epsilon-delta definition of functions that every continuous function is lower semi-continuous as well. (Indeed, a function is both lower and upper semi-continuous if and only if it is continuous. See notes in Appendix A.5.)

The following result follows from the definition of lower semi-continuity.

Theorem 4.27. *Let $A = \{f_\alpha\}$ be a collection of lower semi-continuous functions. Define function g by*

$$g(x) = \sup_{\alpha \in A} f_\alpha(x).$$

Then the function g is lower semi-continuous.

Proof in Appendix A.6.

Theorem 4.28. *(Baire's Theorem on semi-continuous functions) Every lower semi-continuous function can be expressed as the pointwise limit of a sequence of continuous functions.*

Proof in Appendix A.7.

Finally, two general results in measure theory:

Theorem 4.29. *If function f is bounded and measurable over a set of finite measure S, it is integrable over S.*

Proof in Appendix A.8.

Theorem 4.30. *Let $\{f_n\}$ be a pointwise sequence of measurable functions, converging to a function g. Then g is measurable.*

Proof in Appendix A.9.

4.4.2. *Statement.* Finally, we show the crucial statement:

Proposition 4.31. *Let c be continuous and bounded. Then duality holds:*

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c d\pi = \sup_{\psi,\phi} \left(\int_Y \phi d\nu + \int_X \psi d\mu \right)$$

where the supremum is over continuous, bounded functions ψ, ϕ satisfying $\phi(y) + \psi(x) \leq c(x, y)$ for all x, y.

Note the slightly stronger assumption compared to previous parts that c is bounded as well.

One last result on probability measures before we begin:

Theorem 4.32. *Let π be a probability measure on $X \times Y$ with marginals μ, ν on X, Y respectively. Then for $f(x)$ integrable on X,*

$$\int_X f(x) d\mu = \int_{X \times Y} f(x) d\pi.$$

There is a slight abuse of notation here; $f : X \to \mathbb{R}$ should be extended to some $\tilde{f} : X \times Y \to \mathbb{R}$ where $\tilde{f}(x, y) = f(x)$, and similarly for functions on Y, and this should be the function integrated with respect to π, but this adds a lot of notational overhead for little gain in clarity. Nevertheless, we can decompose π into integration with respect to μ, then ν, and as $\nu(Y) = 1$, the two sides are in fact equal.

Onto the proof.

Proof. From Section 4.2, there exists a c-cyclically monotone transference plan π with marginals μ and ν on X, Y respectively, and from Section 4.3, there exists a $\psi : X \to \mathbb{R}$ such that $\Gamma = \text{Support}(\pi) \subset \partial_c \psi$. Now, define ϕ:

$$\phi(y) := \psi^c.$$

We now wish to show that ψ and ϕ are measurable.

Recall the definition of ψ in Equation (4.23):

$$\psi(x) = \sup_{m \in \mathbb{N}} \sup((c(x_1, y_0) - c(x_0, y_0)) + (c(x_2, y_1) - c(x_1, y_1)) + \cdots + (c(x, y_m) - c(x_m, y_m)) : (x_1, y_1), \cdots, (x_m, y_m) \in \Gamma).$$

For a fixed m and a fixed sequence of points $(x_1, y_1), \cdots, (x_m, y_m)$, the function

$$(c(x_1, y_0) - c(x_0, y_0)) + (c(x_2, y_1) - c(x_1, y_1)) + \cdots + (c(x, y_m) - c(x_m, y_m))$$

is the sum of some continuous functions, so it is continuous. If we now define a sequence of functions ψ_m by

$$\psi_m(x) := \sup((c(x_1, y_0) - c(x_0, y_0)) + (c(x_2, y_1) - c(x_1, y_1)) + \cdots + (c(x, y_m) - c(x_m, y_m)) : (x_1, y_1), \cdots, (x_m, y_m) \in \Gamma)$$

as each ψ_m is the supremum over some continuous (and lower semi-continuous) functions, by Theorem 4.27 it must be lower semi-continuous as well. Finally, as

$$\psi(x) = \sup_{m \in \mathbb{N}} \psi_m,$$

ψ must be lower semi-continuous by the same theorem. By Baire's theorem (Theorem 4.28), ψ can be expressed as the pointwise limit of continuous functions, which are measurable, and so by Theorem 4.30, ψ is also measurable.

Now, we examine the measurability of ϕ. We know that

$$\phi(y) = \inf_{x \in X} (-\psi(x) + c(x, y)),$$

and furthermore, note that for fixed x, $-\psi(x) + c(x, y)$ is a continuous function. Then by the same reasoning as above, $\phi(y)$ is a lower semi-continuous function and thus measurable as well.

We examine one more condition necessary for integrability: boundedness. Define

$$\|c\| := \sup_{x,y} c(x, y).$$

We know that $\|c\| < \infty$ as c is bounded. Then, we choose some $(x_0, y_0) \in \partial_c \psi$ such that $-\infty < \psi(x_0) < \infty$. (The ψ so constructed in Section 4.3 does satisfy $\psi(x_0) = 0$ for some x_0, so this is possible.) Then $-\infty < \phi(y_0) < \infty$, as $(x_0, y_0) \in \partial_c \psi$ and for points in $\partial_c \psi$, we have that $\psi(x) + \phi(y) = c(x, y)$.

Then, for any $x \in X$, as $\psi = \phi^c$, we know that

$$\begin{aligned}\psi(x) &= \inf_{y \in Y}(-\phi(y) + c(x,y)) \\ &\leq -\phi(y_0) + c(x, y_0) \\ &\leq -\phi(y_0) + \|c\|\end{aligned}$$

and so ψ is bounded above.
Similarly,

$$\begin{aligned}\phi(y) &= \inf_{x \in X}(-\psi(x) + c(x,y)) \\ &\leq -\psi(x_0) + c(x_0, y) \\ &\leq -\psi(x_0) + \|c\|\end{aligned}$$

and so ϕ is bounded above.

Now, we use the fact that ψ is $c-$concave. We know from its definition that $\psi^c = \phi$, and from Theorem 4.25, we have that $\phi^c = \psi$. Then, we can bound both ψ and ϕ from the other direction:
For all (x, y), $c(x, y) > -\infty$, and $\psi(x) < \infty$. Thus

$$c(x,y) - \psi(x) > -\infty,$$

and taking the infimum over the left-hand side, we have that

$$\phi(y) = \inf_{y \in Y}(c(x,y) - \psi(x)) > -\infty.$$

A similar argument applies to show that $\psi(x) > -\infty$ as well. Thus we have shown that both ψ and ϕ are bounded and measurable, and so from Theorem 4.29 they are integrable.

Finally, as π has marginals μ, ν on X, Y respectively, we apply Theorem 4.32 to $\int_X \psi(x)d\mu$ and $\int_Y \phi(y)d\nu$ to get that

$$\int_Y \phi(y)d\nu - \int_X \psi(x)d\mu = \int_{X\times Y} \phi(y) - \psi(x)d\pi.$$

Recall that Γ, the support of π, satisfies $\Gamma \subset \partial_c\psi$, and so at every $(x, y) \in \Gamma$, we have that

$$\phi(y) - \psi(x) = c(x,y).$$

Thus

$$\int_{X\times Y} \phi(y) - \psi(x)d\pi = \int_{X\times Y} c(x,y)d\pi.$$

Looking back at the statement for weak Monge-Kantorovich as in Equation (3.1), as we have found ψ, ϕ such that equality holds, the infimum of the left-hand side must be equal to the supremum of the right-hand side, and so the duality condition holds. And this is what we wanted to show. □

4.5. **Conclusion.** The main idea of this proof was to produce a π c-cyclically monotone, then to carefully pick a ψ such that ψ and its conjugate ϕ satisfy

$$\psi(x) + \phi(y) = c(x, y)$$

almost everywhere on π. To produce π, we built up the general case from a sequence of discrete cases, and ψ was carefully constructed as to be conveniently c-concave. We can extend this methodology slightly farther to c lower semi-continuous instead of c continuous: Baire's theorem on lower semi-continuous functions (Theorem 4.28) suggests that the solutions for continuous cost functions approximating a lower semi-continuous cost function should approach a solution for the lower semi-continuous cost function as well.

The major implications of this theorem are discussed in Section 2.3.1 above. The economically-minded reader will have noticed that if one inverts all the signs, this problem of minimizing combined costs and maximizing separated prices can actually be seen as a statement on how prices set by a cooperating pair depend on their own costs. Similar manipulations allow this to model how workers and firms are matched, and similar examples abound.

5. Applications to Principal-Agent Problems

5.1. **Overview.** The principal-agent problem consists of two actors, the *principal* and *agent*. The agent has the ability to perform actions not directly observable by the principal, but whose results are observable. Based on what the principal can observe, the principal attempts to design a mechanism that will motivate the agent to behave in a desired fashion. This relates to the subset of game theory called *mechanism design*, and we will demonstrate its connection with the topic of optimal transport in an unusual way.

5.2. **Incentive Design: Toy Example.** To begin with, let's examine an extremely simplified problem in which the power of mechanism design becomes apparent. The following example is adapted from Dixit and Nalebuff's excellent book *Thinking Strategically* [4].

5.2.1. *Problem Statement.* You (the principal) are in charge of a drug manufacture, seeking to bring a new anti-cancer drug to market. Being an established manufacturer, you have the facilities and such necessary for testing, and need only to hire the employees needed to properly test the drug and evaluate its benefits and drawbacks.

The success or failure of your drug hinges on the FDA's decision to approve it for human use. If it is a success, you estimate that it will bring your company a lifetime profit of $400,000 dollars; if it fails, your company receives nothing. The FDA's decision, in turn, hinges on the effort that your lab technicians put in. Your lab technicians can choose to put in high-quality effort: spending late nights in the lab, designing inventive studies, and presenting their findings in neat, well-documented presentations. Or they can put in only a routine effort: eight-hour workdays, one-hour lunch breaks, and all their data scribbled onto a heap of loose-leaf paper.

Obviously, routine effort is easier for the technicians than high-quality effort. For $100,000, you can hire technicians to put out routine effort, but to motivate them to put out high-quality effort, you would need to pay them $140,000. High-quality effort gives your product an 80% chance of success, routine effort only gives your product a 60% chance of success. What is the best course of action?

At this point, high-school mathematics solves this problem. We calculate the expected profit to you (the manufacturer) for routine and high-quality effort:

$$\mathbb{E}(\text{profit} \mid \text{high-quality}) = 80\% \cdot 400 - 140 = 180.$$

$$\mathbb{E}(\text{profit} \mid \text{routine}) = 60\% \cdot 400 - 100 = 140.$$

From this cost-benefit analysis, we see that it is optimal for you to pay higher wages and get lab techs to put forth high-quality effort. But there's a catch: as a pointy-haired middle-manager, you can't evaluate the quality of the lab techs' work. The one-hour lunch breaks might just be productive meetings scheduled between major tests, and the scribbles on notebook paper might be just what the regulators at the FDA are looking for. So if you can't tell the difference between high-quality and routine effort, what stops your workers from demanding high-quality salary and putting forth routine-quality work?

5.2.2. *Solution Mechanism.* The solution to this problem hinges on the one thing you can observe: the success or failure of your product. If you paid your lab techs significantly more if the product succeeds, and less if it fails, they, too, will be motivated to put forth more effort to increase the probability that the product succeeds. But what is the least you can pay to your employees to make this happen? After all, every dollar they don't receive is another dollar of pure profit to you.

Let the amount paid to the employee upon success be S, and the same amount upon failure be F. Then the employee should expect to make at least \$100,000 when using routine effort, and at least \$140,000 when using high-quality effort:

$$\mathbb{E}(\text{wages} \mid \text{high-quality}) = 80\% \cdot S + 20\% \cdot F \geq 140,000$$

$$\mathbb{E}(\text{wages} \mid \text{routine}) = 60\% \cdot S + 40\% \cdot F \geq 100,000.$$

If we solve for equality in both these inequalities, we get that $S = \$180,000$, $F = \$ - 20,000$ - so the employee gets $\$180,000$ if the product succeeds, and must pay the company $\$20,000$ if the product fails. Note that this is equivalent to selling the employee the rights to half the profit in exchange for $\$20,000$ and their labor.

Under this incentive scheme, the employee is now properly motivated to output high-quality labor, and all it costs you is $\$140,000$ - exactly the same as if you could actually observe the quality of labor being produced firsthand. In this example, we can see the power of a properly implemented incentive scheme: even though we can't observe employee effort, we can induce them to act as though we could.

5.3. **Mathematical Footing.** Let's formalize some of the mathematical intuition behind this problem.

5.3.1. *Problem Statement.* Consider the problem of a limited-information monopolist selling a selection of goods to a collection of consumers with differentiated preferences. In other words, the monopolist is the only seller in the market, and consumers can choose to buy only from the monopolist. Consumers will vary in that they derive different amounts of benefit from different types of goods. Most importantly, the monopolist has little or no information about how the consumer values goods, although they may have some beliefs on the distribution of consumer preferences over the entire market.

The monopolist seeks to maximize its own profit - it wants to sell cars to consumers which maximize its own profits, and to this end, using their monopoly position in the market, they can offer a consumer the choice between a single car and no car at all. Most consumers prefer having a car to not having one, but sometimes if the price is too high or the car is too unsuited to their needs, they may choose to reject the car. To address this, the monopolist allows consumers to communicate their preferences in what kind of car they want, and will adjust which car they offer accordingly.

Such markets, or at least markets similar to these, do exist in the real world. For example, the market for cars in an isolated rural area with only one car dealership follows this pattern - consumers come into the dealership with their own wants and desires, and the dealership can effectively name the prices for the different types of

cars that they sell.

5.3.2. *Formalization.* We model this as a mathematical game in several steps. Let Ω be the space of types of customers, and let Y be the space of products which the monopolist sells.
Every customer is an *agent* who possesses a type $\theta \in \Omega$, known only to themselves. This θ encodes all the data about their preferences. These agents then transmit to the principal a "claimed preference" $\tilde{\theta} \in \Omega$, which may or may not be equal to θ. Agents do want to communicate some information about their preferences to the principal in order to try to get a better deal, but giving away too much information could open them up to exploitation.
The monopolist has a variety of products $P \subset Y$ to sell to the consumer. Based on the type $\tilde{\theta}$ announced by the customer, the monopolist offers a product $T(\tilde{\theta})$ for sale to the consumer at a price $v(T(\tilde{\theta}))$.

Definition 5.1. *We label this combination of a matching scheme T and a pricing scheme v a* ***mechanism****.*

Finally, the consumer buys the product $T(\tilde{\theta})$ at a price $v(T(\tilde{\theta}))$. They receive utility equal to some

$$h(\theta, T(\tilde{\theta})) - v(T(\tilde{\theta})).$$

The game moves in a predefined order:

(1) The monopolist announces their mechanism, communicating both their matching and pricing scheme to all customers.
(2) The consumers review their true types θ and communicate to the principal their declared type $\tilde{\theta}$.
(3) The principal communicates to the consumer which product they are allowed to buy, and at what price.
(4) The consumers can choose to buy the product and extract value from it, lessened by the loss in cash of the price of the product.

Given this setup, some natural questions arise: With respect to a given mechanism, how should a consumer behave to maximize their own utility? What mechanism should the monopolist use to maximize their own profit? Is there a mechanism which induces consumers to always reveal their true preferences? We address these questions in order.

5.3.3. *Consumer-side optimization.* This problem is actually rather straightforwards, as the monopolist is forced to announce their mechanism beforehand. For every possible product $y \in Y$, the consumer can consider a set $M_y \subset \Omega$ such that for every $\theta \in M_y$, $t(\theta) = y$. (This is the set of all declared types which are matched with good y.) Then the customer just has to find the y maximizing:

$$\max_{y \in Y} \max_{\theta \in M_y} h(\theta, y) - v(y).$$

The customer figures out which type θ will bring the best price for each product, and how much benefit each product will bring after accounting for this cost. Then, the customer simply chooses the best product.

5.3.4. *Monopolist Profit Maximization.* If we assume that the monopolist has some cost of production $c(y)$ for good y, and the consumer receives some outside utility $u_0(\theta)$ for not buying a car, then we start to wonder how the monopolist should match consumers and price their products to extract maximum value. Let μ be the distribution of consumers over the space Ω, and let $u(\theta)$ be the maximum utility achievable by the consumer of type θ by buying a car, such that

$$u(\theta) := \max_{y \in Y} h(\theta, y) - v(y).$$

For mechanism (T, v) as defined in Definition 5.1, the total profit that the monopolist makes is (assuming truthful reporting by the consumer, which we will address later)

$$\int_\Omega h(\theta, T(\theta)) - v(T(\theta)) d\mu$$

with the additional conditions that $u(\theta) \geq u_0(\theta)$. The monopolist's problem is to find the mechanism (T, v) that maximizes this quantity.

Unfortunately, this is a difficult problem that is not yet solved. The core problem is to demonstrate that the constraints on T, v restrict them to a convex space, and then linear programming techniques may be brought to bear. Additional references are included in the Appendix.

5.3.5. *Incentive Compatibility.* While the problem of maximizing monopolist profits may not be solved, we can better address the problem of inducing all consumers to reveal their true preferences. We formalize this:

Definition 5.2. *Let (T, v) be a mechanism. Then we say that (T, v) is **implementable in dominant strategy** or simply implementable, if for all $(\theta, \tilde{\theta}) \in \Omega^2$:*

$$h(\theta, T(\theta)) - v(T(\theta)) \geq h(\theta, T(\tilde{\theta})) - v(\tilde{\theta}).$$

Roughly speaking, a strategy which is implementable in dominant strategy provides sufficient incentive for every consumer to tell the truth θ about their own preferences, instead of strategically lying with a $\tilde{\theta}$.

Now, we come to the main result of the section: checking if a mechanism is implementable is equivalent to solving an optimal transport problem. We will provide without proof some theorems of Carlier, and explain their relevance to economic theory in general. (References to the papers in question are in Appendix B.2.)

Theorem 5.3. *Let Ω be a bounded connected subset of $\mathbb{R}^n$, μ some probability on Ω absolutely continuous and with a positive Radon-Nikodym derivative with respect to the Lebesgue measure on $\mathbb{R}^n$, and such that $\mu(\partial\Omega) = 0$.*
Let Y be a compact Polish probability space and let ν be a probability measure on Y.
Let $h : \overline{\Omega} \times Y \to \mathbb{R}$ be a continuous function such that:

- *For every ω relatively compact in Ω, there exists a $c(\omega) > 0$ such that for all $(x_1, x_2) \in \omega^2$,*

$$\sup_{y \in Y} |h(x_1, y) - h(x_2, y)| \leq c(\omega)|x_1 - x_2|$$

- *For all $y_0 \in Y$, $h(x, y_0)$ is differentiable in Ω (with respect to x) and for all $(x_0, y_1, y_2) \in \Omega \times Y^2$,*

$$\frac{\partial h}{\partial x}(x_0, y_1) = \frac{\partial h}{\partial x}(x_0, y_2) \implies y_1 = y_2.$$

Then, the Monge problem, the Monge-Kantorovich problem, and the dual problem all admit a solution, the ψ, ϕ maximizing the dual problem satisfy $\psi^h = \phi$, $\phi^h = \psi$, and there exists a Borel map $s : \Omega \to Y$ such that $\psi(x) + \phi(s(x)) = h(x, s(x))$ for all $x \in \Omega$, s solves the Monge problem and (id, s) solves the Monge-Kantorovich problem.

We will not go into details on what this theorem is saying, but note that, given these stronger conditions on the shape of the problem, this theorem allows us to solve the Monge problem as well. This function s is a "choice function" related to the π we were interested in above. However, instead of allowing mass to be split, as in Kantorovich's problem, for every point x in Ω, all the mass of x must go to a single point $s(x)$ in Y.

Theorem 5.4. *Let $s : \Omega \to \mathbb{R}^N$ be a function. (We assume the space of declarable types is a subset of $\mathbb{R}^N$.) Let $h : \overline{\Omega} \times \mathbb{R}^N \to \mathbb{R}$. The following are equivalent:*

- *s is implementable in dominant strategy (as in Definition 5.2).*
- *There exists some $\psi : \Omega \to \mathbb{R}$ which is h-convex and satisfies $\partial^h \psi \neq \emptyset$ such that for all $\theta \in \Omega$,*

$$s(\theta) \in \partial^h \psi(\theta).$$

h-convexity is the natural counterpart to h-concavity, using a supremum instead of an infimum, but it's not essential. But what we really care about is this idea that if h admits a ψ with these nice properties, we can come up with an s which is implementable in dominant strategy. This leads into the next theorem:

Theorem 5.5. *Let $s_0 : \Omega \to Y$ be an arbitrary Borel function. Then there exists a unique Borel map $\overline{s} : \Omega \to Y$ such that:*

- *$\overline{s}$ is implementable in dominant strategy.*
- *s_0 and $\overline{s}$ are equimeasurable: For μ, ν distributions on Ω, Y respectively, $s_{0\#}\mu = \overline{s}_{\#}\mu$ and $s_{0\#}\nu = \overline{s}_{\#}\nu$.*

And $\overline{s}$ is the solution of the Monge problem:

$$\sup_{s \in P(\mu,\nu)} \int_\Omega h(\theta, s(\theta)) d\mu(\theta).$$

This tells us that for any matching s_0, we can come up with a matching $\overline{s}$ which matches the same populations, as the two functions are equimeasurable. Furthermore, $\overline{s}$ is implementable in dominant strategy, which is a significant result.

In this way, this problem of figuring out how to market to consumers can reduce itself down to a question of how to match products to consumers. Then, as this is a matching problem, we can look at it through the framework of optimal transport to gain additional insight into how price shifts move consumption from one product to another.

6. Acknowledgements

My thanks to Phillip Lo, my REU mentor, who provided plenty of feedback and helpful ideas to help get this project off the ground. Additionally, I owe a great deal to Peter May, whose tireless efforts ensure that we have a high-quality REU program every year. Thanks also are due to professor Greg Lawler, whose thought-provoking lunches and excellent lectures on probability and analysis fed my interest and gave me the mathematical intuition to grapple with these problems. Shoutouts to the entire Honors Analysis gang as well.

7. Bibliography

References

[1] Patrick Billingsley. Convergence in Probability Measures. John Wiley & Sons, Inc. 1999
[2] Guillaume Carlier. Duality and existence for a class of mass transportation problems and economic applications. In: Kusuoka, S., Maruyama, T. (eds) Advances in Mathematical Economics. Advances in Mathematical Economics, vol 5. Springer, Tokyo. https://doi.org/10.1007/978-4-431-53979-7_1.
[3] Guillaume Carlier. Nonparametric Adverse Selection Problems. Annals of Operations Research 114, 71–82 (2002). https://doi.org/10.1023/A:1021001917492
[4] Avinash K. Dixit and Barry J. Nalebuff. Thinking Strategically. W. W. Norton & Company. 1991
[5] Rick Durrett. Probability: Theory and Examples. https://services.math.duke.edu/~rtd/PTE/PTE5_011119.pdf.
[6] Alessio Figalli, Young-Heon Kim, Robert J. McCann. "When is multidimensional screening a convex program?". Journal of Economic Theory, Volume 146, Issue 2, 2011, Pages 454-478, ISSN 0022-0531, https://doi.org/10.1016/j.jet.2010.11.006. (https://www.sciencedirect.com/science/article/pii/S0022053111000093)
[7] Matematleta (https://math.stackexchange.com/users/138929/matematleta), Lower Semicontinuous Function = Supremum of Sequence of Continuous Functions, URL (version: 2019-12-29): https://math.stackexchange.com/q/3490563
[8] MÉRIGOT, QUENTIN, and ÉDOUARD OUDET. "HANDLING CONVEXITY-LIKE CONSTRAINTS IN VARIATIONAL PROBLEMS." SIAM Journal on Numerical Analysis 52, no. 5 (2014): 2466–87. http://www.jstor.org/stable/24512202.
[9] H. L. Royden, P. M. Fitzpatrick. Real Analysis. 2010.
[10] Walter Rudin. Real and Complex Analysis. McGraw-Hill, Inc. 1987
[11] Chapter 11. Convergence in Distribution. https://sites.stat.washington.edu/jaw/COURSES/520s/522/HO.522.20/ch11c.pdf
[12] V. S. Varadarajan. "On the Convergence of Sample Probability Distributions." Sankhyā: The Indian Journal of Statistics (1933-1960) 19, no. 1/2 (1958): 23–26. http://www.jstor.org/stable/25048365.
[13] Alfred Galichon. Optimal Transport Methods in Economics. Princeton University Press. 2016.
[14] Cedric Villani. Optimal Transport, old and new. Springer. 2008.

. Appendix

I will attempt to provide proofs or links to proofs that are as clear and convincing as possible.

Appendix A. Supplementary Theorems

A.1. **Portmanteau Theorem.** The Portmanteau Theorem demonstrates that the following conditions are equivalent:

(1) A sequence of measures $\{\mu_n\}$ on X converges weakly to measure μ.

(2) $\lim_{n\to\infty} \int_X f d\mu_n = \int_X f d\mu$ for all bounded continuous functions f.
(3) $\lim_{n\to\infty} \int_X f d\mu_n = \int_X f d\mu$ for all bounded Lipschitz functions f.
(4) $\limsup_{n\to\infty} \int_X f d\mu_n \leq \int_X f d\mu$ for all upper semi-continuous functions f bounded above.
(5) $\liminf_{n\to\infty} \int_X f d\mu_n \geq \int_X f d\mu$ for all lower semi-continuous functions f bounded below.
(6) $\limsup_{n\to\infty} \mu_n(C) \leq \mu(C)$ for all closed sets C.
(7) $\liminf_{n\to\infty} \mu_n(U) \geq \mu(U)$ for all open sets U.

Billingsley [1] demonstrates that 1, 2, 6, and 7 are equivalent, and is a good place to start. Note that Lipschitz functions are continuous, so 3 follows from 2. This pdf [11] (which I believe is from a Washington University STAT 522 class) demonstrates how 4 and 5 follow from 3, and eventually 6 and 7 follow.

A.2. **Law of Large Numbers for Empirical Measures.** This theorem is equivalent to the Glivenko-Cantelli Theorem. Durrett provides a proof on p. 79 of [5].

A.3. **Prokhorov's Theorem.** Billingsley [1] devotes all of section 5 in his book to Prokhorov's theorem, corollaries, and explanation. It's quite readable.

A.4. **Fenchel Transform of c-concave Function.** We can prove this from the definitions.

Proof. First, we show that for any function $\psi : X \to \mathbb{R}$, $\psi^{ccc} = \psi^c$. First, note that

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

Expanding out $\psi^c(y)$, this is

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

We can distribute the sign on the inside and pull the inner infimum out to get

$$\psi^{cc}(x) = \inf_{y\in Y}\sup_{\tilde{x}\in X}(c(x,y) - c(\tilde{x},y) + \psi(\tilde{x})).$$

Repeating the same process for ψ^{ccc}, we have that

$$\psi^{ccc}(y) = \inf_{x\in X}\sup_{\tilde{y}\in Y}\inf_{\tilde{x}\in X}(c(x,y) - c(x,\tilde{y}) + c(\tilde{x},\tilde{y}) - \psi(\tilde{x})).$$

Considering the case where $\tilde{x} = x$, this expression simplifies to

$$\psi^{ccc}(y) = \inf_{x\in X}\sup_{\tilde{y}\in Y}(c(x,y) - \psi(x))$$

which is equal to $\psi^c(y)$, and so $\psi^c(y) \geq \psi^{ccc}(y)$. Considering when $\tilde{y} = y$, a similar line of reasoning shows that $\psi^{ccc}(y) \geq \psi^c(y)$, so we ultimately have that $\psi^{ccc}(y) = \psi^c(y)$.

Then, if ψ is c-concave, there exists a ϕ such that

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

so $\psi = \phi^c$. Then $\psi^{cc} = \phi^{ccc} = \phi^c$, so $\psi = \psi^{cc}$ when ψ is c-concave. □

The Fenchel conjugate is quite interesting. This particular theorem is closely related to the Fenchel-Moreau Theorem, which provides both sufficient and necessary conditions for $\psi^{cc} = \psi$.

A.5. **Lower and upper semi-continuity.** We can show that a function which is both lower and upper semi-continuous is continuous directly from the epsilon-important definition of continuity.

Proof. Fix x_0, and take an $\epsilon > 0$. Then by lower and upper semi-continuity, there exist neighborhoods U and V of x_0 such that for $x \in U \cap V$, $f(x_0) - \epsilon < f(x) < f(x_0) + \epsilon$. $U \cap V$ is a neighborhood of x_0, so there exists an ball of radius δ within $U \cap V$ centered at x_0 for some $\delta > 0$. As we can find such a δ for every ϵ, f is continuous.

Suppose f is continuous. Let $\epsilon = |y - f(x_0)|$. Then by continuity, there exists $\delta > 0$ such that for x satisfying $|x_0 - x| < \delta$, $|f(x_0) - f(x)| < \epsilon$. Thus $f(x_0) - \epsilon < f(x) < f(x_0) + \epsilon$. For $y > f(x_0)$, we have $f(x) < y$, and for $y < f(x_0)$, we have $f(x) > y$. Thus we can show that f is both lower and upper semi-continuous. □

A.6. **Supremum of lower semi-continuous functions.** I will provide an actual proof, as this is fairly direct.

Proof. For a point x_0, fix a $y < g(x_0)$. Then, as g is defined as the pointwise supremum of the collection of lower semi-continuous functions A, there must exist some f_i such that $f_i(x_0) > \frac{f(x_0)+y}{2}$. But as f_i is lower semi-continuous, there must exist some neighborhood U of x_0 such that for $x \in U$, $f_i(x) > y$. Finally, $g(x) \geq f_i(x)$ for all x, so for all $x \in U$, $g(x) > y$ as well, and so g is lower semi-continuous at x_0. The same argument applies to every point. □

A.7. **Baire's Theorem.** Baire proved this result in 1905, but his paper is in French. The only other textbook I could find that covers this theorem is excessively general, so I'll cite this Math Stack Exchange post [7] instead. The main idea is that for $F(x)$ lower semi-continuous that we want to show is the pointwise supremum of continuous functions, we consider the functions

$$F_n(x) = \inf_{y \in X} (F(y) + n \cdot d(x, y)),$$

where d is the distance function, and we show that the F_n are increasing in n, continuous, and converge to F.

A.8. **Integrability of Bounded and Measurable Functions.** This is sometimes taken as a definition, using simple functions to approximate measurable functions. Royden and Fitzpatrick prove this as a theorem in [9].

A.9. **Convergence of Measurable Functions.** This is shown in Rudin's *Real and Complex Analysis* [10], a classical reference.

Appendix B. Further Readings

The material in this paper was taken from a broad variety of sources. Cedric Villani's textbooks on optimal transport are a good starting place. His *Optimal Transport: old and new* is especially valuable, as it provides complete proofs and extensive commentary on related readings.

B.1. **Monopolist Profit Maximizing.** Figalli, Kim, and McCann ([6]) provide necessary conditions for h so that h-convex functions are convex. Merigot and Oudet tackle this problem numerically in [8]. However, much work remains to be done. The generalized Spence-Mirrlees condition appears to be the main focus of research.

B.2. **Incentive Compatibility Theorems.** These results are derived from a series of papers by Guilliame Carlier. Theorem 5.4 is from [3]. Theorem 5.3 and Theorem 5.5 are from [2].