Online lectures on optimal transport and applications to quantile regression

Alfred Galichon (NYU Econ+Courant and Sciences Po)

Invited lectures at HU Berlin
November 18-19, 2020

- These two online lectures will provide a mathematical and computational introduction to optimal transport, in connection with quantile regression.
- Code in R is provided in the following Github directory (`https://github.com/alfredgalichon/Humboldt-2020`). Other language work but the LP solver used will be Gurobi, so the language of choice should have a convenient interface to these.
- Outline:
  D1: optimal transport, quantiles, vector quantiles
  D2: conditional quantiles, quantile regression, vector quantile regression

Texts on optimal transport

- Mathematical foundations:
  - [OTON] C. Villani, *Optimal Transport: Old and New*, AMS, 2008.
  - [OTAM] F. Santambrogio, *Optimal Transport for Applied Mathematicians*, Birkhäuser, 2015.
- Fluid mechanics point of view:
  - [TOT] C. Villani, *Topics in Optimal Transportation*, AMS, 2003.
- Computational focus:
  - [NOT] G. Peyré, M. Cuturi (2018). *Numerical optimal transport*, Arxiv.
- Economics focus:
  - [OTME] A. Galichon. *Optimal Transport Methods in Economics*, Princeton, 2016.

Lecture 1. The Monge-Kantorovich duality: general overview and linear programming

Refs: [OTME], Chapters 1, 2 and 8.
Complement: [TOT], Chapter 1.

Section 1

The Monge-Kantorovich theorem

- Consider the problem of assigning a possibly infinite number of workers and firms. Each worker should work for one firm, and each firm should hire one worker.
- Workers and firms have heterogenous characteristics; let $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ be the vector of characteristics of workers and firms respectively. For all practical purpose in this course, they will be either open or closed subsets of $\mathbb{R}^d$.
- Workers and firms are in equal mass, which is normalized to one. The distribution of worker's types is P, and the distribution of the firm's types is Q, where P and Q are probability measures on $\mathcal{X}$ and $\mathcal{Y}$.

- A *coupling* determines which workers are assigned to which firms. If we had a finite number of workers and firms, we would need to count the number of workers of a given type matched with firms of a given type. More generally, a coupling will be defined as the probability measure π of occurence of worker-firm pairs. If $(X, Y) \sim \pi$ is the joint random pair, then $X \sim P$ and $Y \sim Q$, where $X \sim P$ means "X has distribution P." In other words, the first *margin* of π should be P, while its second margin should be Q.
- This motivates the following definition:

Definition

The set of couplings of probability distributions P and Q is the set of probability distributions over $\mathcal{X} \times \mathcal{Y}$ with first and second margins P and Q. This set is denoted $\mathcal{M}(P, Q)$. That is, a probability measure π over $\mathcal{X} \times \mathcal{Y}$ is in $\mathcal{M}(P, Q)$ if and only if

$$\pi(A \times \mathcal{Y}) = P(A) \text{ and } \pi(\mathcal{X} \times B) = Q(B)$$

holds for every subset A of $\mathcal{X}$ and B of $\mathcal{Y}$. By extension, a random pair $(X, Y) \sim \pi$ where $\pi \in \mathcal{M}(P, Q)$ will also be called a coupling of P and Q.

- *Independent coupling*, a.ka. *random matching*: $\pi(A, B) = P(A) Q(B)$, so that, if $(X, Y) \sim \pi$, then $X \sim P$ and $Y \sim Q$ are independent.
- *Pure assignment*, or *Monge coupling*: Y is a deterministic function of X; that is, $Y = T(X)$. In our worker-firm example, this assumes that every workers of type x will get assigned the same type of firm, $T(x) \in \mathcal{Y}$. Then if $X \sim P$, then $T(X) \sim Q$, which we denote by

$$T\#P = Q \tag{1}$$

  where $T\#P$ is the distribution of $T(X)$ when $X \sim P$ ("push-forward" of P by map T, sometimes denoted $PT^{-1} = Q$).
- In general, a coupling is associated to a *Markov kernels*, $\pi(dy|x)$ such that

$$\int_{\mathcal{X}} \pi(B|x)\, dP(x) = Q(B)$$

  for every subset B of $\mathcal{Y}$.

- Assume that if worker x works for firm y, this generates a quantity of output $\Phi(x, y)$, measured in some monetary unit. A social planner decides which workers to assign to which firms and seeks to maximize the total output. The theory Optimal Transport studies how to do this.
- The *Monge Problem* (posed at the end of the 18th century) consists of looking among all the *pure* assignments, that is,

$$\max_{T(.)} \mathbb{E}_P \left[\Phi\left(X, T\left(X\right)\right)\right]. \tag{2}$$
$$s.t.\ T\#P = Q$$

- In general, the Monge problem is difficult:
  - It is nonlinear. In the discrete case, assuming P and Q have N equally weighted sample points, this amounts to looking for T within the permutations of $\{1, ..., N\}$: $N!$ possibilities.
  - It may not have a solution. Think e.g. of the discrete case when P has less sample points than Q.
  - In fact, it remained unsolved for more than a century!

- Kantorovich in the 1940s came up with the idea of *linear programming Relaxation*: instead of looking among Monge couplings, look among all the couplings. Hence, instead of maximizing $\mathbb{E}_P [\Phi (X, T(X))]$ s.t. $T\#P = Q$, simply maximize $\mathbb{E}_\pi [\Phi (X, Y)]$ s.t. $X \sim P$ and $Y \sim Q$.
- This leads to the *Kantorovich problem*

$$\max_{\pi \in \mathcal{M}(P,Q)} \mathbb{E}_\pi [\Phi (X, Y)] \tag{3}$$

- This formulation has many advantages:
  - it is a linear programming problem, albeit an infinite-dimensional one. The dual is very informative.
  - It has a solution π under very weak assumptions.
  - In a number of relevant cases, the solution to the Monge and Kantorovich problems will coincide.

Theorem

Let $\mathcal{X}$ and $\mathcal{Y}$ be two Banach spaces, and let P and Q be two probability measures on $\mathcal{X}$ and $\mathcal{Y}$ respectively. Let $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ be an upper semicontinuous surplus function bounded above by $\overline{a}(x) + \overline{b}(y)$ where $\overline{a}$ and $\overline{b}$ are respectively integrable with respect to P and Q. Then: (i) The value of the primal Monge-Kantorovich problem

$$\sup_{\pi \in \mathcal{M}(P,Q)} \mathbb{E}_\pi \left[\Phi(X, Y) \right] \tag{4}$$

coincides with the value of the dual

$$\begin{aligned} &\inf_{u,v} \mathbb{E}_P \left[u(X) \right] + \mathbb{E}_Q \left[v(Y) \right] \\ &s.t.\ u(x) + v(y) \geq \Phi(x, y) \end{aligned} \tag{5}$$

where the infimum is over measurable and integrable functions u and v, and the inequality constraint should be satisfied for almost every x and almost every y (all these statements are respective to measures P and Q).
(ii) An optimal solution π to problem (4) exists.

Theorem

(iii) Assume further Φ is bounded below by $\underline{a}(x) + \underline{b}(y)$ where $\underline{a}$ and $\underline{b}$ are respectively integrable with respect to P and Q. Then the dual problem (5) also has solutions.

The value of the primal problem $\sup_{\pi \in \mathcal{M}(P,Q)} \mathbb{E}_\pi \left[\Phi \left(X, Y \right) \right]$ can be rewritten

$$\sup_{\pi \in \mathcal{M}^+} \int \Phi \left(x, y \right) d\pi \left(x, y \right) + A_{P,Q} \left(\pi \right),$$

where $\mathcal{M}^+$ is the set of positive measures over $\mathcal{X} \times \mathcal{Y}$ (not necessarily of total mass one, and not necessarily with fixed marginals), and $A_{P,Q}$ should be such that

$$A_{P,Q} \left(\pi \right) = \begin{cases} 0 \text{ if } \pi \in \mathcal{M} \left(P, Q \right) \\ = -\infty \text{ else.} \end{cases}$$

One can take

$$A_{P,Q} \left(\pi \right) = \inf_{u,v} \int u \left(x \right) dP \left(x \right) + \int v \left(y \right) dQ \left(y \right) - \int \left(u \left(x \right) + v \left(y \right) \right) d\pi \left(x, y \right)$$

so that the value of the primal problem becomes

$$\sup_{\pi \in \mathcal{M}^+} \inf_{u,v} \left\{ \begin{array}{c} \int \Phi \left(x, y \right) - \left(u \left(x \right) + v \left(y \right) \right) d\pi \left(x, y \right) \\ + \int u \left(x \right) dP \left(x \right) + \int v \left(y \right) dQ \left(y \right) \end{array} \right\}.$$

It is the case here that sup inf = inf sup (this fact will be admitted without a proof), which yields

$$\inf_{u,v} \int u(x)\, dP(x) + \int v(y)\, dQ(y) + B_\Phi(u,v)$$

where $B_\Phi(u,v) = \sup_{\pi \in \mathcal{M}^+} \int \Phi(x,y) - (u(x) + v(y))\, d\pi(x,y)$, so that

$$B_\Phi(u,v) = \begin{cases} 0 \text{ if } u(x) + v(y) \geq \Phi(x,y) \text{ for all } x \text{ and } y \\ = +\infty \text{ else.} \end{cases}$$

thus the value of the problem rewrites as (5).This argument is only a rough sketch; the minmax principle which we invoked when inverting the sup and the inf needs to be carefully established, and the spaces in which the functions u and v and the measure μ live need to be made precise. See Villani's [TOT], Ch. 1 for a rigorous argument.

We argue that $u(x)$ can be interpreted as the equilibrium wage of worker x, while $v(y)$ can be interpreted as the equilibrium profit of firm y.

Proposition

If (u, v) is a solution to the dual of the Kantorovich problem, then we can always redefine u and v so that they take value $+\infty$ outside of the supports of P and Q, respectively. In this case,

$$u(x) = \sup_y (\Phi(x, y) - v(y)) \quad (6)$$

$$v(y) = \sup_x (\Phi(x, y) - u(x)) \quad (7)$$

should hold almost surely with respect to the probabilities P and Q, respectively.

Section 2

Discrete case and linear programming

- Assume that the type spaces $\mathcal{X}$ and $\mathcal{Y}$ are finite, so $\mathcal{X} = \{1, ..., N\}$, and $\mathcal{Y} = \{1, ..., M\}$.
- The total mass of workers and jobs is normalized to one. The mass of workers of type x is p_x; the mass of jobs of type y is q_y, with $\sum_x p_x = \sum_y q_y = 1$.
- Let π_{xy} be the mass of workers of type x assigned to jobs of type y. Every worker is occupied and every job is filled, thus

$$\sum_{y=1}^{M} \pi_{xy} = p_x \text{ and } \sum_{x=1}^{N} \pi_{xy} = q_y. \tag{8}$$

(Note that this formulation does not restrict to Monge couplings, i.e. it allows for $\pi_{xy} > 0$ and $\pi_{xy'} > 0$ to hold simultaneously with $y \neq y'$.)

- Assume the economic output created when assigning worker x to job y is Φ_{xy}. Hence, under assignment π, the total output is $\sum_{xy} \pi_{xy} \Phi_{xy}$.
- Thus, the optimal assignment is

$$\max_{\pi \geq 0} \sum_{xy} \pi_{xy} \Phi_{xy} \tag{9}$$

$$s.t. \sum_{y=1}^{M} \pi_{xy} = p_x \ \ [u_x]$$

$$\sum_{x=1}^{N} \pi_{xy} = q_y \ \ [v_y]$$

and it is now a finite-dimensional linear programming problem.

- Note that it is nothing else than the Monge-Kantorovich problem when P and Q are discrete probability measures on $\mathcal{X} = \{1, ..., N\}$, and $\mathcal{Y} = \{1, ..., M\}$.

Theorem (M-K, finite-dimensional case)

(i) The value of the primal problem (9) coincides with the value of the dual problem

$$\min_{u,v} \sum_{x=1}^{N} p_x u_x + \sum_{y=1}^{M} q_y v_y. \tag{10}$$

$$s.t. u_x + v_y \geq \Phi_{xy} \ \ [\pi_{xy} \geq 0]$$

(ii) Both the primal and the dual problems have optimal solutions. If π is a solution to the primal problem and (u, v) a solution to the dual problem, then by complementary slackness,

$$\pi_{xy} > 0 \text{ implies } u_x + v_y = \Phi_{xy}. \tag{11}$$

(iii) If (u, v) is a solution to the dual problem, then

$$u_x = \max_{y \in \{1,...,M\}} \left\{ \Phi_{xy} - v_y \right\} \text{ and } v_y = \max_{x \in \{1,...,N\}} \left\{ \Phi_{xy} - u_x \right\}. \tag{12}$$

- This problem of computation of the Optimal Assignment Problem, more specifically of (π, u, v), is arguably the most studied problem in Computer Science, and dozens, if not hundreds of algorithms exist, whose running time is polynomial in $\max(n, m)$, typically a power less than three of the latter.
- Famous algorithms include: the Hungarian algorithm (Kuhn-Munkres); Bertsekas' auction algorithm; Goldberg and Kennedy's pseudoflow algorithm. See an introduction to these algorithms in http://www.assignmentproblems.com/doc/LSAPIntroduction.pdf.
- Here, we will show how to solve the problem with the help of a linear programming solver used as a black box; our challenge here will be to carefully set up the constraint matrix as a sparse matrix in order to let a large scale linear programming solvers such as Gurobi recognize and exploit the sparsity of the problem.

- Let Π and Φ be the matrices with typical elements (π_{xy}) and (Φ_{xy}). We let p, q, u, v, and 1 the column vectors with entries (p_x), (q_y), (u_x), (v_y), and 1, respectively. Problem (9) rewrites using matrix algebra as

$$\max_{\Pi \geq 0} Tr\left(\Pi' \Phi\right) \tag{13}$$
$$\Pi 1_M = p$$
$$1'_N \Pi = q'.$$

- We need to convert matrices into vectors; this can be done for instance by stacking the columns of Π into a single column vector (typical in R or Matlab). This operation is called *vectorization*, which we will denote

$$vec\left(A\right),$$

which reshapes a $N \times M$ matrix into a $nm \times 1$ vector. In R, this is implemented by `c(A)`; in Matlab, by `reshape(A,[n*m,1])`.

- The objective function rewrites as

$$vec\left(\Phi\right)' vec\left(\Pi\right).$$

- Recall that if A is a $M \times p$ matrix and B a $N \times q$ matrix, then the Kronecker product $A \otimes B$ of A and B is a $mn \times pq$ matrix such that

$$vec\left(BXA'\right) = \left(A \otimes B\right) vec\left(X\right). \tag{14}$$

In R, $A \otimes B$ is implemented by `kronecker(A,B)`; in Matlab, by `kron(A,B)`.

- The first constraint $I_N \Pi 1_M = p$, vectorizes therefore as

$$\left(1_M' \otimes I_N\right) vec\left(\Pi\right) = vec\left(p\right),$$

and similarly, the second constraint $1_N' \Pi I_M = q'$, vectorizes as

$$\left(I_M \otimes 1_N'\right) vec\left(\Pi\right) = vec\left(q\right).$$

- Note that the matrix $1_M' \otimes I_N$ is of size $N \times NM$, and the matrix $I_M \otimes 1_N'$ is of size $M \times NM$; hence the full matrix involved in the left-hand side of the constraints is of size $\left(N + M\right) \times NM$. In spite of its large size, this matrix is *sparse*. In R, the identity matrix I_N is coded as `sparseMatrix(1:N,1:N)`, in Matlab as `Speye(N)`.

- Setting $z = vec\,(\Pi)$, the linear programming problem then becomes

$$\begin{aligned} &\max_{z\geq 0} vec\,(\Phi)' z \\ s.t.\ &\left(1'_M \otimes I_N\right) z = vec\,(p) \\ &\left(I_M \otimes 1'_N\right) z = vec\,(q') \end{aligned} \tag{15}$$

which is ready to be passed on to a linear programming solver such as Gurobi.

- A LP solver typically computes programs of the form

$$\begin{aligned} &\max_{z\geq 0} c' z \\ &s.t.\ Az = d. \end{aligned} \tag{16}$$

In R, Gurobi is called to compute program (16) by `gurobi(list(A=A,obj=c,modelsense="max",rhs=d,sense="="))`.

Programming example: optimal assignment via Gurobi

See subdirectory `PE1a-OptimalAssignment/` in the Github repository.

Section 3

Univariate case and quantiles

- Assume $\mathcal{X} = \mathcal{Y} = \mathbb{R}$. This implies that both workers and firms are characterized by scalar characteristics x and y. Restrictive, but most of the applied literature in Economics to this day has focused on this case, which already generates interesting economic insights.
- Assume that there is the same number of workers and firms, and this number is normalized to one. Let P be the probability distribution of the workers types x, and Q the probability distribution of the firms types y.
- E.g. the literature on CEO compensation studies the matching problem of firms and managers: each manager has a measure of talent $x \in \mathbb{R}_+$ (extra return generated), and each firm has market capitalization $y \in \mathbb{R}_+$. Then, the economic value generated by a manager of talent x running a firm of size y is $\Phi(x, y) = xy$. Assume manager of talent x is assigned to firm of size $y = T(x)$. The constraint on the assignment map T is that $T\#P = Q$, which means that each firm is run by a manager and each manager runs a firm. Then the total value created is

$$\mathbb{E}_P[\Phi(X, T(X))] = \mathbb{E}[XT(X)].$$

- Under a natural assuption on Φ, called *supermodularity*, we will see that the optimal coupling is such that $Y = T(X)$, where T is nondecreasing. T will be given an explicit characterization.

- Recall that given a probability distribution P on the real line, one can define the *quantile* of that distribution, denoted F_P^{-1}, which is a map from $[0,1]$ to R, nondecreasing and continuous from the right, and such that if $U \sim \mathcal{U} := \mathcal{U}([0,1])$, then

$$F_P^{-1}(U) \sim P. \tag{17}$$

  F_P^{-1} is the generalized inverse of the c.d.f. of P, F_P, the proper inverse when P has a positive density, in which case $U = F_P(X)$. Note also that if $X \sim P$, then $F_P(X)$ has a uniform distribution if and only if P has no mass point.

- Representation (17) extends to the case of bivariate distributions: for $\pi \in \mathcal{M}(P,Q)$, there exists a pair (U,V) of uniform random variables such that

$$\left(F_P^{-1}(U), F_Q^{-1}(V)\right) \sim \pi, \tag{18}$$

  and the c.d.f. associated with the distribution of (U,V) is called the *copula* associated with distribution π.

- A pair of random variables (X, Y) is *comonotone* if there is $U \sim \mathcal{U}$ such that $X = F_P^{-1}(U)$ and $Y = F_Q^{-1}(U)$. Equivalently, X and Y are said to exhibit *Positive Assortative Matching (PAM)*.
- The copula associated with a pair of comonotone random variables is the c.d.f. associated with (U, U), which is $F(u, v) = \min(u, v)$. This copula is called the *upper Fréchet-Hoeffding copula*.
- Note that when the cdf of X is continuous, there is a much simpler equivalent statement of comonotonicity:

Lemma

If the distribution of X has no mass points, then X and Y are comonotone if and only if there exists a nondecreasing map T such that $Y = T(X)$. Moreover, one can choose $T(x) = F_Q^{-1}(F_P(x))$.

Proof.

Consider $U \sim \mathcal{U}$ such that $X = F_P^{-1}(U)$ and $Y = F_Q^{-1}(U)$. If the distribution of X has no mass point, then $U = F_P(X)$. Hence, $Y = F_Q^{-1}(F_P(X))$. □

- Assume Φ is *supermodular*, that is, for every scalars x, x', y and y',

$$\Phi\left(x \vee x', y \vee y'\right) + \Phi\left(x \wedge x', y \wedge y'\right) \geq \Phi\left(x, y\right) + \Phi\left(x', y'\right), \qquad (19)$$

where $x \vee x'$ and $x \wedge x'$ denote respectively the maximum and the minimum between scalars x and x'. When Φ is twice continuously differentiable (which we will assume from now on), this is equivalent to

$$\frac{\partial^2 \Phi\left(x, y\right)}{\partial x \partial y} \geq 0. \qquad (20)$$

- Assume that there are two types of workers $\mathcal{X} = \{\underline{x}, \bar{x}\}$ and firms $\mathcal{Y} = \{\underline{y}, \bar{y}\}$. An equivalent restatement of Condition (19) is then

$$\bar{x} \geq \underline{x} \text{ and } \bar{y} \geq \underline{y} \text{ implies } \Phi\left(\bar{x}, \bar{y}\right) + \Phi\left(\underline{x}, \underline{y}\right) \geq \Phi\left(\bar{x}, \underline{y}\right) + \Phi\left(\underline{x}, \bar{y}\right) \qquad (21)$$

which asserts that the total output created is higher if the high types match together and the low types match together (assortative matching) rather than if mixed high/low pairs are formed.

Example

The following examples of surplus functions are supermodular:
(i) Cobb-Douglas function: $\Phi(x, y) = x^a y^b$ $(x, y \geq 0)$, with $a, b \geq 0$,
(ii) General multiplicative form: $\Phi(x, y) = \zeta(x)\,\xi(y)$ with ζ and ξ nondecreasing,
(iii) Leontieff: $\Phi(x, y) = \min(x, y)$,
(iv) C.E.S. function: $\Phi(x, y) = \left(x^{-\rho} + y^{-\rho}\right)^{-1/\rho}$, $\rho \geq 0$,
(v) $\Phi(x, y) = \phi(x - y)$ where ϕ is concave; in particular, $\Phi(x, y) = -|x - y|^p$, $p \geq 1$ or $\Phi(x, y) = -(x - y - k)^+$,
(vi) $\Phi(x, y) = \phi(x + y)$, where ϕ convex.

Theorem

(i) Assume that Φ is supermodular. Then the primal of the Monge-Kantorovich problem

$$\sup_{\pi \in \mathcal{M}(P,Q)} \mathbb{E}_{\pi} \left[\Phi \left(X, Y \right) \right] \tag{22}$$

has a comonotone solution.
(ii) Conversely, if Problem (22) has a comonotone solution for any choice of probability distributions P and Q on the real line, then Φ is supermodular.
(iii) If, in addition, P has no mass points, then there is an optimal assignment which is is pure and satisfies $Y = T(X)$ where

$$T(x) = F_Q^{-1} \circ F_P(x). \tag{23}$$

The proof of part (i) is based on the following lemma.

Lemma

Let Z_1 and Z_2 be two Bernoulli random variables of respective success probability p_1 and p_2. Then $\mathbb{E}[Z_1 Z_2] \leq \min(p_1, p_2)$.

Proof.

As $Z_2 \leq 1$, $\mathbb{E}[Z_1 Z_2] \leq \mathbb{E}[Z_1] = p_1$. Similarly $\mathbb{E}[Z_1 Z_2] \leq \mathbb{E}[Z_2] = p_2$.
Thus, $\mathbb{E}[Z_1 Z_2] \leq \min(p_1, p_2)$. □

We are now ready to sketch the proof of Theorem 8.

Sketch of proof of Theorem 8.

(i) Take $U \sim \mathcal{U}$, and $X = F_P^{-1}(U)$ and $Y = F_Q^{-1}(U)$. By (17), $X \sim P$ and $Y \sim Q$ and (X, Y) is comonotone by definition. The proof is in three steps.
Step 1. For $a, b \in \mathbb{R}$, consider surplus function $\phi_{ab}(x, y) := 1\{x \geq a\} 1\{y \geq b\}$, and let $Z_1 = 1\{X \geq a\}$ and $Z_2 = 1\{Y \geq b\}$. Z_1 and Z_2 are two Bernoulli random variables of respective success probability $p_1 = 1 - F_P(a)$ and $p_2 = 1 - F_Q(b)$, thus $\mathbb{E}[Z_1 Z_2] \leq \min(p_1, p_2)$, but a straightforward calculation shows that the inequality actually holds as an equality. Hence (X, Y), which is comonotone, is optimal for each surplus function ϕ_{ab}.
Step 2. Assume $\mathcal{X} = [\underline{x}, \bar{x}]$ and $\mathcal{Y} = [\underline{y}, \bar{y}]$ are compact intervals. Then

$$F(x, y) = \frac{\Phi(x, y) - \Phi(\underline{x}, y) - \Phi(x, \underline{y}) + \Phi(\underline{x}, \underline{y})}{\Phi(\bar{x}, \bar{y}) - \Phi(\underline{x}, \bar{y}) - \Phi(\bar{x}, \underline{y}) + \Phi(\underline{x}, \underline{y})}$$

is a c.d.f. associated to a probability measure ζ, and hence $F(x, y) = \iint \phi_{ab}(x, y)\, d\zeta(a, b)$. □

Proof.

As a result, if $\pi \in \mathcal{M}(p, q)$ is the distribution of (X, Y) where X and Y are comonotone, then

$$\int F(x, y)\, d\pi(x, y) \geq \int F(x, y)\, d\tilde{\pi}(x, y)$$

for every $\tilde{\pi} \in \mathcal{M}(p, q)$. But as F is of the form $F(x, y) = K\Phi(x, y) + f(x) + g(y) + c$ with $K > 0$, and because $\int \{f(x) + g(y) + c\}\, d\pi(x, y) = \int \{f(x) + g(y) + c\}\, d\tilde{\pi}(x, y)$ for every $\tilde{\pi} \in \mathcal{M}(p, q)$, it results that

$$\int \Phi(x, y)\, d\pi(x, y) \geq \int \Phi(x, y)\, d\tilde{\pi}(x, y) \;\; \forall \tilde{\pi} \in \mathcal{M}(p, q)$$

which completes step 2. $\square$

Proof.

Step 3. When $\mathcal{X}$ and $\mathcal{Y}$ are the real line, the result still holds by approximation.

(ii) The converse follows by taking for P the discrete probability with two mass points $\underline{x}$ and $\bar{x}$ with probability $1/2$ each, and Q the discrete probability with two mass points $\underline{y}$ and $\bar{y}$ also each with probability $1/2$. Then if (22) has a solution such that $F_P^{-1}(U)$ and $Y = F_Q^{-1}(U)$, for $U \sim \mathcal{U}([0,1])$, it follows that condition (21) holds.

(iii) follows from (i) and Lemma 6. □

Note that the assumptions made in Theorem 8 do not guarantee that all the optimal assignments are comonotone. Indeed, the trivial example where $\Phi(x, y) = 0$ for every x and y provides an example of supermodular surplus function, for which any assignment is optimal. For this reason, we provide a strengthening of the previous result, which ensures uniqueness. We will assume Φ is strictly supermodular, that is if both $\bar{x} > \underline{x}$ and $\bar{y} > \underline{y}$ hold, then $\Phi(\bar{x}, \bar{y}) + \Phi(\underline{x}, \underline{y}) > \Phi(\bar{x}, \underline{y}) + \Phi(\underline{x}, \bar{y})$.

Theorem

Assume that Φ is strictly supermodular, and P has no mass point. Then the primal Monge-Kantorovich problem (22) has a unique optimal assignment, and this assignment is characterized by $Y = T(X)$ where T is given by (23).

- Assume (u, v) is a solution to the dual of the Monge-Kantorovich problem

$$\begin{aligned} &\inf \mathbb{E}_P\left[u\left(X\right)\right] + \mathbb{E}_Q\left[v\left(Y\right)\right] \\ &s.t.\ u\left(x\right) + v\left(y\right) \geq \Phi\left(x, y\right) \end{aligned} \tag{24}$$

- Then $v(y)$ is intepreted as the value of the problem of a firm of type y, choosing the optimal worker x. Then the firm's program is

$$v\left(y\right) = \max_x \left\{\Phi\left(x, y\right) - u\left(x\right)\right\}$$

thus by first order conditions, one is led to the *wage equation*

$$u'\left(x\right) = \frac{\partial \Phi}{\partial x}\left(x, T\left(x\right)\right), \tag{25}$$

where T is given by (23).

Theorem

(i) Assume Φ is supermodular and continuously differentiable with respect to its first variable. Assume P has no mass point. Then the dual Monge-Kantorovich problem (24) has a solution (u, v). Further, u solves the wage equation (25). Hence, u is determined up to a constant c by

$$u(x) = c + \int_{x_0}^{x} \frac{\partial \Phi}{\partial x}(t, T(t))\, dt. \tag{26}$$

(ii) Assume further that Q has no mass point, and that Φ is also continuously differentiable with respect to its second variable. Then v is given by

$$v(y) = c' + \int_{T(x_0)}^{y} \frac{\partial \Phi}{\partial y}\left(T^{-1}(z), z\right) dz, \tag{27}$$

where c and c' are related by $c + c' = \Phi(x_0, T(x_0))$.

As a consequence of the previous considerations, we have the following result:

Theorem

Assume $\mu = \mathcal{U}([0,1])$ is the uniform distribution on the $[0,1]$ interval, and Q is a probability distribution on the real line. Consider the optimal transport problem

$$\sup_{\pi \in \mathcal{M}(\mu, Q)} \mathbb{E}_\pi [UY] = \inf_{\varphi(u) \text{ convex lsc}} \int_0^1 \varphi(u)\, du + \int \varphi^*(y)\, dQ(y)$$

then a solution of the prial problem is given by $\left(U, F_Q^{-1}(U)\right)$, while a solution of the dual problem is given by

$$\varphi(u) = \int_0^u F_Q^{-1}(t)\, dt.$$

Section 4

Vector quantiles

Theorem (Brenier)

Assume that μ and ν have finite second moments, and μ has a density. Then the solution $(U, Y) \sim \pi \in \mathcal{M}(\mu, \nu)$ to the primal problem is represented by

$$Y = \nabla \varphi (U)$$

where (φ, φ^*) is a solution to the dual problem. Such φ is unique up to a constant.

Intuition of the proof: if φ is differentiable, then y is matched with u that maximizes $\{u^\intercal y - \varphi(u)\}$ over $u \in \mathbb{R}^d$. By first order conditions, such u satisfy $\nabla \varphi (u) = y$. It turns out, however, that differentiability is not a serious concern (at least, almost never).

The previous result allows to provide a representation of a large class of probability distributions ν over $\mathbb{R}^d$ as the probability distribution of $\nabla \varphi \left(U \right)$, for U with a fixed distribution μ. There is however a limitation, in the sense that it requires that ν has finite second moments, which is needed for the optimal transport problem to have a finite value. Fortunately, McCann's theorem addresses this issue:

Theorem (McCann)

Assume that μ and ν are probability distributions such that μ has a density. Then there is a unique (up to a constant) function u such that

$$Y = \nabla \varphi \left(U \right)$$

holds almost surely with $U \sim \mu$ and $Y \sim \nu$.

The latter is denoted $\nabla \varphi \# \mu = \nu$.

- Definition. Let μ be a distribution on $\mathbb{R}^d$. For every distibution ν on $\mathbb{R}^d$, the μ-vector quantile associated with distribution ν is the map

$$Q_\nu(u) = \nabla \varphi(u)$$

where φ is a convex function such that $\nabla \varphi \# \mu = \nu$.

- The existence and uniqueness of this object are provided by McCann's theorem.
- We will take $\mu = \mathcal{U}([0,1]^d)$. In that case, when $\nu = \nu_1 \otimes \nu_2 \otimes ... \otimes \nu_d$, we have

$$\varphi(u) = \sum_{i=1}^{d} \varphi_i(u_i),$$

so $Q_\nu(u) = \left(\varphi_1'(u_1), \varphi_2'(u_2), ..., \varphi_d'(u_d)\right) = \left(F_{\nu_1}^{-1}(u_1), F_{\nu_2}^{-1}(u_2), ..., F_{\nu_d}^{-1}(u_d)\right)$.