

When an inf is also a sup: Fenchel-Rockafellar and Kantorovich duality

Grad Analysis Seminar, 2/6/12

Abstract: Kantorovich duality says that the least amount of effort required to rearrange one pile of dirt to look like another is equal to the largest amount an enterprising friend could charge to perform the task for you (even with some constraints on his pricing structure). Thinking of probability measures as piles of dirt, the least amount of effort required to make one look like another provides a way to measure their distance, and Kantorovich duality tells you how to approximate this distance from above and below. I will reformulate the (discrete) Kantorovich problem as the minimization of a sum of two convex functions and prove Kantorovich duality as a consequence of Fenchel-Rockafellar.

I created these notes as a rough guideline to use during my talk—apologies for any type-o’s they contain!

1 Big picture: why it’s useful to know when an inf is also a sup

Imagine you can write some quantity of interest as

$$\text{quantity of interest} = \inf\{x : x \in A\}$$

If you can reformulate this as a sup

$$\sup\{y : y \in B\} = \text{quantity of interest} = \inf\{x : x \in A\},$$

then you can approximate it from both above and below, i.e. for $x \in A, y \in B$,

$$y \leq \sup\{y : y \in B\} = \text{quantity of interest} = \inf\{x : x \in A\} \leq x.$$

The sup problem is dual to your original inf problem. Today, we’re going to show that the discrete Kantorovich problem can be cast as either an inf or a sup, which will in turn tell us how to approximate the distance between probability measures from above and below.

2 The Discrete Kantorovich Problem

Consider two probability measures μ and ν on a discrete set $X = \{1, \dots, n\}$. They assign some non-negative mass to each point, so that when you sum over all the points, you get one. [Draw pictures of μ and ν .] (The case of probability measures on \mathbb{R}^n is similar, but the proofs are quicker and the main idea comes through better in the discrete case.)

We can define vectors $\vec{\mu}, \vec{\nu} \in \mathbb{R}^n$ that say how much mass is at each point: $\vec{\mu}_i = \mu(\{i\})$.

Think of μ and ν as two ways of piling dirt. The Kantorovich problem is: what is the least amount of effort that it takes to rearrange the dirt in μ to look like ν ?

This might seem like an artificial problem, but it turns out that you can measure the distance between probability measures based on the amount of effort it takes to rearrange one to look like the other. This metric is a really useful way to think about distance in the context of diffusive PDEs, especially when it comes to constructing approximate solutions.

To write down the Kantorovich problem rigorously, first we need some notation:

Notation: Let E be the set of $n \times n$ real valued matrices, E_+ be the set of matrices with nonnegative entries, and $\vec{1} = \{1, \dots, 1\}$. E is a Hilbert space with the inner product $\langle A, B \rangle = \text{Tr}(A^t B) = \sum_i \sum_j A_{ij} B_{ij}$ and Hilbert Schmidt norm $\|A\| = (\text{Tr}(A^t A))^{1/2} = (\sum_i \sum_j A_{ij}^2)^{1/2}$ (can think of matrices as vectors of length n^2 and the inner product as the dot product). If \vec{v} and \vec{w} are column vectors, the inner product is $\vec{v} \cdot \vec{w} = w^t v = v_1 w_1 + \dots + v_n w_n$ and the outer product is

$$\vec{v} \otimes \vec{w} = \vec{v} \vec{w}^t = \begin{pmatrix} v_1 w_1 & v_1 w_2 & \dots & v_1 w_n \\ v_2 w_1 & v_2 w_2 & \dots & v_2 w_n \\ \dots & \dots & \dots & \dots \\ v_n w_1 & v_n w_2 & \dots & v_n w_n \end{pmatrix}$$

Note that

$$\vec{v} \otimes \vec{u} = \begin{pmatrix} v_1 & v_1 & \dots & v_1 \\ v_2 & v_2 & \dots & v_2 \\ \dots & \dots & \dots & \dots \\ v_n & v_n & \dots & v_n \end{pmatrix}.$$

I will call this a ‘‘column vector matrix.’’ Likewise,

$$\vec{u} \otimes \vec{w} = \begin{pmatrix} w_1 & w_2 & \dots & w_n \\ w_1 & w_2 & \dots & w_n \\ \dots & \dots & \dots & \dots \\ w_1 & w_2 & \dots & w_n \end{pmatrix}.$$

I will call this a ‘‘row vector matrix.’’

A transportation plan is a way of rearranging the dirt in μ to look like ν . We can write a transportation plan as

$X \in E_+ : X_{ij} =$ the amount of mass you need to move from point i to point j

$$K(\mu, \nu) = \{\text{admissible transportation plans}\} = \{X \in E_+ | X \vec{u} = \mu, X^t \vec{u} = \nu\}$$

Finally, we’re interested in moving dirt with the least effort, but how much effort does it take to move dirt from a point i to a point j ? We can specify this with a cost function

$C \in E_+ : C_{ij} =$ the amount of effort to move a unit of dirt from a point i to a point j

The minimal amount of effort to rearrange μ to look like ν is:

$$\inf\{\text{Tr}(C^t X) : X \in K(\mu, \nu)\} = \inf\left\{\sum_{i=1}^n \sum_{j=1}^n C_{ij} X_{ij} : X \in K(\mu, \nu)\right\}$$

Now imagine that you're getting ready to go all this effort to rearrange the pile of dirt μ to look like the pile of dirt ν , and suddenly a friend comes up and offers you a proposition: he has a wheel barrow and will happily move all the dirt—for a price.

$$w_i = \text{amount per unit of dirt from } i$$

$$z_j = \text{amount per unit of dirt to } j$$

(Think of these as vectors \vec{w} and \vec{z} .)

Since he is anxious for your business, he promises you that he won't charge you more to move the dirt from point i to point j than it would cost you in terms of effort:

$$\vec{w}_i + \vec{z}_j \leq C_{ij} \quad \forall i, j,$$

or writing things out as matrices,

$$C - \vec{w} \otimes \vec{u} - \vec{u} \otimes \vec{z} \in E_+.$$

The total amount that you pay him will be

$$\vec{\mu} \cdot \vec{w} + \vec{\nu} \cdot \vec{z}.$$

Naturally, within the limits of his promise, your friend wants to charge you as much as he can. The most he can charge is:

$$\sup\{\vec{\mu} \cdot \vec{w} + \vec{\nu} \cdot \vec{z} : C - \vec{w} \otimes \vec{u} - \vec{u} \otimes \vec{z} \in E_+\}$$

Because of his promise, the amount he will charge will always be less than the amount of effort you would expend:

$$\inf\{\text{Tr}(C^t X) : X \in K(\mu, \nu)\} \geq \sup\{\vec{\mu} \cdot \vec{w} + \vec{\nu} \cdot \vec{z} : C - \vec{w} \otimes \vec{u} - \vec{u} \otimes \vec{z} \in E_+\}$$

Kantorovich duality says that they are actually equal:

$$\inf\{\text{Tr}(C^t X) : X \in K(\mu, \nu)\} = \sup\{\vec{\mu} \cdot \vec{w} + \vec{\nu} \cdot \vec{z} : C - \vec{w} \otimes \vec{u} - \vec{u} \otimes \vec{z} \in E_+\}$$

Aside from knowing when your friend is giving you a good deal versus a mediocre deal, Kantorovich duality is extremely useful when the minimal effort to transport the dirt represents the distance between two probability measures: Kantorovich duality gives you a way to estimate the distance from above and below.

- Pick any $X \in K(\mu, \nu)$, the amount of effort will be greater than the distance

$$\text{Tr}(C^t X) \geq \inf\{\text{Tr}(C^t X) : X \in K(\mu, \nu)\} = d(\mu, \nu)$$

- Pick any pricing structure w_i, z_j your friend could have - the amount he will charge will be less than the distance:

$$d(\mu, \nu) = \sup\{\vec{\mu} \cdot \vec{w} + \vec{\nu} \cdot \vec{z} : C - \vec{w} \otimes \vec{u} - \vec{u} \otimes \vec{z} \in E_+\} \geq \vec{\mu} \cdot \vec{w} + \vec{\nu} \cdot \vec{z}$$

The proof of Kantorovich duality follows from some facts about convex functions and the Fenchel-Rockafellar theorem, which tells you how to reformulate many types of “infs” as “sups”.

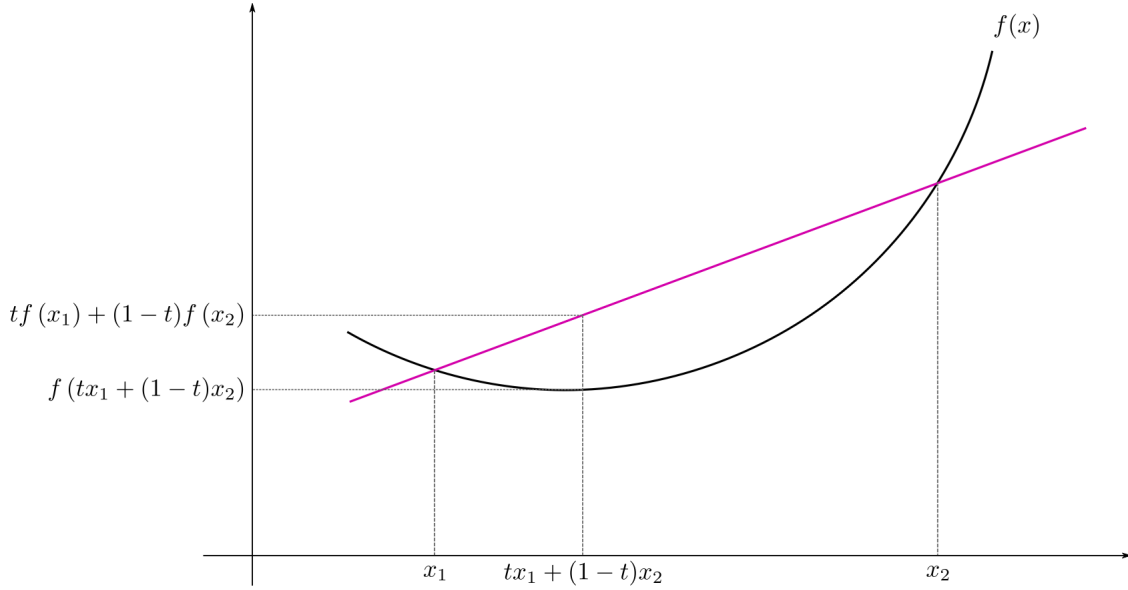
3 Background: Convex Analysis

Let E be a real vector space. A subset $F \subseteq E$ is **convex** if $\forall X, Y \in F, t \in [0, 1]$,

$$(1 - t)X + tY \in E. \quad \text{“convex combination”}$$

A function $\varphi : E \rightarrow \mathbb{R} \cup \{\infty\}$ is **convex** if for all $x, y \in E, t \in [0, 1]$,

$$\varphi((1 - t)x + ty) \leq (1 - t)\varphi(x) + t\varphi(y).$$



The **domain** of φ is the set $\text{Dom}(\varphi) = \{x : \varphi(x) < \infty\}$. φ is **proper** if $\text{Dom}(\varphi) \neq \emptyset$.

Given $\varphi : E \rightarrow \mathbb{R} \cup \{\infty\}$, we may define the conjugate function $\varphi^* : E^* \rightarrow \mathbb{R} \cup \{\infty\}$ by

$$\varphi^*(f) := \sup_{x \in E} \{ \langle f, x \rangle - \varphi(x) \}.$$

Fenchel-Rockafellar Theorem. *Let φ and ψ be two proper convex functions on E . Suppose there exists some $x_0 \in \text{Dom}(\varphi) \cap \text{Dom}(\psi)$ at which φ or ψ is continuous. Then*

$$\inf \{ \varphi(x) + \psi(x) : x \in E \} = \sup \{ -\varphi^*(f) - \psi^*(-f) : f \in E^* \}.$$

4 Proof of Discrete Monge-Kantorovich Duality

Kantorovich Duality Theorem.

$$\inf \{ \text{Tr}(C^t X) : X \in K(\mu, \nu) \} = \sup \{ \vec{\mu} \cdot \vec{w} + \vec{\nu} \cdot \vec{z} : C - \vec{w} \otimes \vec{u} - \vec{u} \otimes \vec{z} \in E_+ \}$$

Proof. We're actually going to apply Fenchel-Rockafellar a little backwards (because it's easier to compute the conjugate functions). Multiplying both sides by -1 , it's enough to show

$$\sup\{-\text{Tr}(C^t X) : X \in K(\mu, \nu)\} = \inf\{-\vec{\mu} \cdot \vec{w} - \vec{\nu} \cdot \vec{z} : C - \vec{w} \otimes \vec{u} - \vec{u} \otimes \vec{z} \in E_+\}.$$

Now, we recast this in the form of Fenchel-Rockafellar by defining two convex functions on E , the Hilbert space of $n \times n$ matrices. First, define φ as:

$$\varphi(X) = \begin{cases} 0 & \text{if } C - X \in E_+ \\ +\infty & \text{otherwise.} \end{cases}$$

We need to show $\forall X, Y \in E, t \in (0, 1)$,

$$\varphi((1-t)X + tY) \leq (1-t)\varphi(X) + t\varphi(Y).$$

- If either X, Y does not satisfy $C - X \in E_+$, the RHS is infinity, and we're done.
- If X and Y are both in the set $\{X \in E : C - X \in E_+\}$, we need to show that any convex combination also is. In other words, we need to show the set is convex. Since,

$$\{X \in E : C - X \in E_+\} = \cap_{i,j} \{X \in E : C_{ij} \geq X_{ij}\},$$

thinking of X as a vector in \mathbb{R}^{n^2} , this is just the intersection of n^2 half spaces, hence a rectangle in \mathbb{R}^{n^2} , hence convex.

Now, define ψ as:

$$\psi(X) = \begin{cases} -\vec{\mu} \cdot \vec{w} - \vec{\nu} \cdot \vec{z} & \text{if } X = \vec{w} \otimes \vec{u} + \vec{u} \otimes \vec{z} \text{ for some } \vec{w}, \vec{z} \in \mathbb{R}^n \\ +\infty & \text{otherwise.} \end{cases}$$

Note that $X = \vec{w} \otimes \vec{u} + \vec{u} \otimes \vec{z}$ means that X is the sum of a column-vector matrix and a row-vector matrix. We need to show $\forall X, Y \in E, t \in (0, 1)$,

$$\psi((1-t)X + tY) \leq (1-t)\psi(X) + t\psi(Y).$$

- If either X or Y is not column vector matrix plus row vector matrix, the RHS is infinity, and we're done.
- If both X, Y are column vector matrix plus row vector matrix, so is their convex combination with column vector $(t\vec{w}_x + (1-t)\vec{w}_y)$ and row vector $((t\vec{z}_x + (1-t)\vec{z}_y)$. Therefore the LHS equals the RHS.

With these definitions of φ and ψ ,

$$\inf\{-\vec{\mu} \cdot \vec{w} - \vec{\nu} \cdot \vec{z} : C - \vec{w} \otimes \vec{u} - \vec{u} \otimes \vec{z} \in E_+\} = \inf\{\varphi(X) + \psi(X) : X \in E\}.$$

To apply Fenchel-Rockafellar, we need to find $X_0 \in \text{Dom}(\varphi) \cap \text{Dom}(\psi)$ at which φ or ψ is continuous. Fix $a < C_{ij}$ and let $X_a = a\vec{u} \otimes \vec{u}$ be the matrix of all a 's. Then

$$\psi(X_a) = -\vec{\mu} \cdot a\vec{u} = -a < \infty,$$

so $X_a \in \text{Dom}(\psi)$. Likewise $\varphi(X_a) = 0$ since $C - X_a \in E_+$, so $X_a \in \text{Dom}(\varphi)$.

Now, note that if X within ϵ of X_a in the Hilbert Schmidt norm, none of the entries of X can be more than ϵ greater than the corresponding entry of X_a . So if we choose ϵ so that $a + \epsilon < C_{ij}$, so that $C - X \in E_+$ for all X within an ϵ ball of X_a . Therefore, φ is identically zero on an ϵ neighborhood of X_a , so φ is continuous at X_a .

Therefore, the Fenchel-Rockafellar theorem guarantees that

$$\inf\{\varphi(X) + \psi(X) : X \in E\} = \sup\{-\varphi^*(-Y) - \psi^*(Y) : Y \in E^*\}.$$

It remains to show the RHS equals what we want. We do this by computing φ^* and ψ^* :

$$\begin{aligned} \psi^*(Y) &= \sup\{\text{Tr}(Y^t X) - \psi(X) : X \in E\} \\ &= \sup\{\text{Tr}(Y^t(\vec{w} \otimes \vec{u} + \vec{u} \otimes \vec{z})) + \vec{\mu} \cdot \vec{w} + \vec{\nu} \cdot \vec{z} : \vec{w}, \vec{z} \in \mathbb{R}^n\} \\ &= \sup\{Y\vec{u} \cdot \vec{w} + Y^t\vec{u} \cdot \vec{z} + \vec{\mu} \cdot \vec{w} + \vec{\nu} \cdot \vec{z} : \vec{w}, \vec{z} \in \mathbb{R}^n\} \\ &= \sup\{(Y\vec{u} + \vec{\mu}) \cdot \vec{w} + (Y^t\vec{u} + \vec{\nu}) \cdot \vec{z} : \vec{w}, \vec{z} \in \mathbb{R}^n\} \\ &= \begin{cases} 0 & \text{if } Y\vec{u} = -\vec{\mu} \text{ and } Y^t\vec{u} = -\nu \\ +\infty & \text{otherwise.} \end{cases} \end{aligned}$$

$$\begin{aligned} \varphi^*(Y) &= \sup\{\text{Tr}(Y^t X) - \varphi(X) : X \in E\} \\ &= \sup\{\text{Tr}(Y^t X) : C - X \in E_+\} \\ &= \begin{cases} \text{Tr}(Y^t C) & \text{if } Y \in E_+ \\ +\infty & \text{otherwise.} \end{cases} \end{aligned}$$

Therefore

$$\begin{aligned} -\varphi^*(Y) - \psi^*(-Y) &= \begin{cases} -\text{Tr}(Y^t C) - 0 & \text{if } Y \in E_+ \text{ and } -Y\vec{u} = -\vec{\mu}, -Y^t\vec{u} = -\nu \\ -\infty & \text{otherwise.} \end{cases} \\ &= \begin{cases} -\text{Tr}(Y^t C) & \text{if } Y \in K(\mu, \nu) \\ -\infty & \text{otherwise.} \end{cases} \end{aligned}$$

Consequently, the Fenchel-Rockafellar result implies

$$\begin{aligned} \inf\{\varphi(X) + \psi(X) : X \in E\} &= \sup\{-\varphi^*(-Y) - \psi^*(Y) : Y \in E^*\} \\ \inf\{-\vec{\mu} \cdot \vec{w} - \vec{\nu} \cdot \vec{z} : C - \vec{w} \otimes \vec{u} - \vec{u} \otimes \vec{z} \in E_+\} &= \sup\{-\text{Tr}(Y^t C) : Y \in K(\mu, \nu)\} \\ \sup\{\vec{\mu} \cdot \vec{w} + \vec{\nu} \cdot \vec{z} : C - \vec{w} \otimes \vec{u} - \vec{u} \otimes \vec{z} \in E_+\} &= \inf\{\text{Tr}(Y^t C) : Y \in K(\mu, \nu)\} \\ \sup\{\vec{\mu} \cdot \vec{w} + \vec{\nu} \cdot \vec{z} : C - \vec{w} \otimes \vec{u} - \vec{u} \otimes \vec{z} \in E_+\} &= \inf\{\text{Tr}(C^t Y) : Y \in K(\mu, \nu)\} \end{aligned}$$

□