1 Review

1.1 Optimal Transport Problem

Recall Kantorovich's optimal transport problem:

Minimize

$$I[\pi] = \int_{X \times Y} c(x,y) d\pi(x,y)$$

where π is a coupling in $\Pi(\mu, \nu)$, c is the cost function. Recall as well that μ and ν are marginals of π with respect to integration over x and y respectively. The optimal transport cost is the minimum value of $I[\pi]$ over couplings of μ and ν.

Also recall that Monge's problem is very similar to the Kantorovich problem, with an additional constraint on π: We must have that

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

for some function T mapping $X \to Y$. To ensure marginalization holds, T must also satisfy

$$\int_X \psi(T(x)) d\mu(x) = \int_Y \psi(y) d\nu(y)$$

for all functions ψ in $L^1(d\nu)$ (or in $L^\infty(d\nu)$). This is essentially restricting the problem to a smaller set of couplings, with the requirement that $\nu = T\#\mu$ ($\nu(B) = \mu(T^{-1}(B))$).

Thus, the Monge problem is: Minimize

$$I[T] = \int_X c(x, T(x)) d\mu(x)$$

where $T\#\mu = \nu$.

1.2 Kantorovich Duality

Now let's recall the statement of Kantorovich duality from last time:

Theorem 1.1. *Kantorovich duality.*
We have spaces X, Y and associated distributions μ, ν, and c is a lower semi-continuous cost function. Let π be a coupling of μ and ν and $(\varphi, \psi) \in L^1(d\mu) \times L^1(d\nu)$. Let

$$I[\pi] = \int_{X \times Y} c(x,y) d\pi(x,y)$$

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

Now, define Φ_c to be all (φ, ψ) from before which also satisfy

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

for $d\mu$-almost all $x \in X$, and $d\nu$-almost all $y \in Y$. Then the following duality statement holds:

$$\inf_{\pi \in \Pi(\mu,\nu)} I[\pi] = \sup_{(\varphi,\psi)\in\Phi_c} J(\varphi,\psi)$$

We can also think of φ, ψ as bounded continuous functions without changing anything.

2 Identification of Optimal Transport with Quadratic Cost

Today we would like to identify when a transport plan is optimal, and also when there are unique transfer plans. We will focus on quadratic costs. It turns out that:

(a) A transfer plan is optimal iff it is concentrated on the subdifferential of a convex function (Knott-Smith).

(b) A transfer plan is unique under some weak regularity condition (Brenier's theorem).

We will see that in particular, if μ and ν are absolutely continuous probability measures, there is a unique mapping $x \to \nabla\varphi(x)$ with φ convex that transports μ onto ν.

We will first show a duality-based approach to showing these results, and then an alternative approach based on the concept of cyclical monotonicity.

Here are the key theorems, which hold under the following assumptions.

(a) Let μ, ν be probability measures on $\mathbb{R}^n$ with finite second order moments:

$$M_2 = \int_{\mathbb{R}^n} \frac{\|x\|_2^2}{2} d\mu(x) + \int_{\mathbb{R}^n} \frac{\|y\|_2^2}{2} d\nu(y)$$

(b) Let the cost be defined $c(x,y) = \frac{1}{2}\|x-y\|_2^2$.

Theorem 2.1. *Knott-Smith Optimality.*
$\pi \in \Pi(\mu,\nu)$ is optimal iff there exists a convex lower semi-continuous function φ such that for $d\pi$-almost all (x,y), $y \in \partial\varphi(x)$ where this is the subdifferential set of φ. Then, (φ, φ^*) is a minimizer of

$$\int_{\mathbb{R}^n} \varphi(x) d\mu(x) + \int_{\mathbb{R}^n} \psi(y) d\nu(y)$$

where $\langle x, y\rangle \le \varphi(x) + \psi(y)$ for all x, y.

Theorem 2.2. *Brenier's Theorem.*
Assume μ is sufficiently nice (doesn't give mass to small sets), then there is a unique optimal π such that

$$d\pi(x,y) = d\mu(x)\mathbf{1}(y = \nabla\varphi(x))$$

where $\nabla\varphi(x)$ is the unique ($d\mu$-almost everywhere) gradient of a convex function such that $\nabla\varphi \# \mu = \nu$ (e.g., $T(x) = \nabla\varphi(x)$ in the Monge problem). Furthermore, if ν is similarly well-behaved, then $\nabla\varphi^*$ is the inverse of $\nabla\varphi$ in both directions, and is the $d\nu$-almost everywhere unique solution of the Monge problem for transporting ν onto μ.

2.1 Duality Reduction

First, let's think about the dual problem. We have that

$$\varphi(x) + \psi(y) \leq \frac{\|x-y\|_2^2}{2}$$

by the definition of Φ_c, and re-arranging, we get

$$\langle x, y\rangle \leq \left(\frac{\|x\|_2^2}{2} - \varphi(x)\right) + \left(\frac{\|y\|_2^2}{2} - \psi(y)\right)$$

We recognize these objects are related to convex conjugate functions, and replace

$$\bar{\varphi}(x) = \frac{\|x\|_2^2}{2} - \varphi(x)$$

$$\bar{\psi}(y) = \frac{\|y\|_2^2}{2} - \psi(y)$$

Then, applying the moment condition, we can write

$$\inf_{\pi\in\Pi(\mu,\nu)} I[\pi] = M_2 - \sup_{\pi\in\Pi(\mu,\nu)} (\int \langle x, y\rangle d\pi(x,y))$$

and likewise

$$\sup_{(\varphi,\psi)\in\Phi_c} J = M_2 - \inf_{\bar{\varphi},\bar{\psi}\in\bar{\Phi}} J(\bar{\varphi},\bar{\psi})$$

where $\bar{\Phi}$ is the set of pairs such that $\langle x, y\rangle \leq \bar{\varphi}(x) + \bar{\psi}(y)$. This yields another form of the duality principle:

$$\inf_{\pi\in\Pi(\mu,\nu)} \int \langle x, y\rangle d\pi(x,y) = \sup_{(\bar{\varphi},\bar{\psi})\in\bar{\Phi}} J(\bar{\varphi},\bar{\psi})$$

Note that solutions to this transformed problem can be transformed into the original problem by calculating

$$\left(\frac{\|x\|_2^2}{2} - \bar{\varphi}(x), \frac{\|y\|_2^2}{2} - \bar{\psi}(y)\right)$$

So we can consider these problems instead. This formulation will be useful in the following proofs.

2.2 Convexity

2.2.1 Convex Conjugates

Before moving on, let's recall what convex conjugate functions are.

Definition 2.3. Convex conjugate.
The Legendre transform of a function φ is

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

Note that for all $x, y \in \mathbb{R}^n$, this also means

$$\langle x, y \rangle \leq \varphi(x) + \varphi^*(y)$$

Some additional properties: If φ is convex lower semi-continuous, then there exists a dual, and also the double dual is the original function. These things all imply each other.

We also present a quick application of the Legendre dual to reduce the number of pairs $(\bar{\varphi}, \bar{\psi}) \in \bar{\Phi}$ that we have to consider.

Lemma 2.4. *Double convexification.*

$$\bar{\psi}(y) \geq \sup_x \{\langle x, y \rangle - \bar{\varphi}(x)\} =: \bar{\varphi}^*(y)$$
$$\bar{\varphi}(x) \geq \sup_y \{\langle x, y \rangle - \bar{\varphi}^*(y)\} =: \varphi^{**}(x)$$

Note that this further implies the following chain:

$$\begin{aligned} J(\bar{\varphi}, \bar{\psi}) &\geq J(\bar{\varphi}, \bar{\varphi}^*) \\ J(\bar{\varphi}, \bar{\varphi}^*) &\geq J(\bar{\varphi}^{**}, \bar{\varphi}^*) \\ \inf_{(\bar{\varphi}, \bar{\psi}) \in \bar{\Phi}} J(\bar{\varphi}, \bar{\psi}) &\geq \inf_{\varphi} J(\varphi^{**}, \varphi^*) \end{aligned} \tag{1}$$

Thus, the infimum is unchanged when you restrict $\bar{\Phi}$ to pairs $(\varphi^{**}, \varphi^*)$, which happen to be convex lower semicontinuous functions.

2.2.2 Gradient Properties

Let's also recall that for φ convex and differentiable, for all $z \in \mathbb{R}^n$,

$$\varphi(z) \geq \varphi(x) + \nabla\varphi(x) \cdot (z - x)$$

and also that $\nabla\varphi$ is monotone:

$$\langle \nabla\varphi(z) - \nabla\varphi(x), z - x \rangle \geq 0$$

2.2.3 Subdifferentials

Subdifferentials generalizes the notion of derivative for a convex function φ. It is a set valued object:

$$\partial\varphi(x) = \{y : \forall z \in \mathbb{R}^n, \varphi(z) \geq \varphi(x) + \langle y, z - x\rangle\}$$

A function is differentiable at a point iff the subgradient contains one element. Additionally, the following characterization will be useful:

Lemma 2.5. *Subdifferential characterization.*
For convex lower semi-continuous φ, for all $x, y \in \mathbb{R}^n$,

$$\langle x, y\rangle = \varphi(x) + \varphi^*(y)$$

is equivalent to $y \in \partial\varphi(x)$ as well as equivalent to $x \in \partial\varphi^*(y)$. The last two are equivalent as well.

2.3 Proof of Knott-Smith Optimality Criterion

We'll forget the bar notation from now on for simplicity.

First we prove the optimality criterion. From last time, we know there is an optimal transport plan π. We now need to prove both directions of the characterization.

We now want to show that if π is optimal, it is concentrated on a subdifferential of a convex function. We then have the following proposition:

Lemma 2.6. *Optimal pair of convex conjugate functions.*
There exists a pair (φ, φ^*) of lower semi-continuous convex conjugate functions such that

$$\inf_{\bar{\Phi}} J = J(\varphi, \varphi^*)$$

Proof. We'll take this for granted. Basically, you need to consider a minimizing sequence (φ_k, ψ_k) and by double convexification assume that they're pairs of convex conjugates. Then we want to show that in the limit, these things are in $L^1(d\mu) \times L^1(d\nu)$ and also that J of the limited versions is $\leq$ the lim inf of J of the sequence versions (this will work by convergence in supremum norm). This is basically true because you can uniformly bound the sequence iterates and satisfy uniform Lipschitz bounds, and thus they converge uniformly. Then double convexify the limited versions to finish. To prove this you also need to ensure that the sequence terms stay finite when the second moment M_2 is finite. □

Thus we now have an optimal pair (φ, φ^*) which are convex lower semi-continuous. By Kantorovich duality and by the fact that π is a coupling, we have

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \langle x, y\rangle d\pi(x, y) = \int_{\mathbb{R}^n} \varphi d\mu + \int_{\mathbb{R}^n} \varphi^* d\nu$$

$$= \int_{\mathbb{R}^n \times \mathbb{R}^n} [\varphi(x) + \varphi^*(y)] d\pi(x, y)$$

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} [\varphi(x) + \varphi^*(y) - \langle x, y \rangle] d\pi(x, y) = 0$$

We know that by the definition of convex conjugate the integrand is nonzero, so it must vanish $d\pi$-almost everywhere. By the characterization of subdifferential, this implies that $y \in \partial\varphi(x)$.

Now we show the other direction: If $y \in \partial\varphi(x)$ $d\pi$-almost everywhere, we need to show that it is an optimal coupling. We can now just reverse the arguments to get

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \langle x, y \rangle d\pi(x, y) = \int_{\mathbb{R}^n} \varphi d\mu + \int_{\mathbb{R}^n} \varphi^* d\nu$$

Thus we proved iff and are done.

2.4 Proof of Brenier's Theorem

Now we want to show for good μ, there is a unique optimal π defined

$$d\pi(x, y) = d\mu(x)\mathbf{1}(y = \nabla\varphi(x))$$

where $\nabla\varphi$ is the unique ($d\mu$-almost everywhere) gradient of convex function s.t. $\nabla\varphi \# \mu = \nu$.

First, by convexity and since $\varphi \in L^1(d\mu)$, all the mass is concentrated on the interior of the domain. Here, negligible part of the set is nondifferentiable. Thus for $d\mu$-almost all x, the subdifferential of φ at x is $\{\nabla\varphi(x)\}$. This is also the case for $d\pi$ (since μ is marginal defined over X). Thus, by the Knott-Smith optimality criterion, $y = \nabla\varphi(x)$ for $d\pi$-almost all (x, y) (since the subdifferential is a singleton at the gradient).

Now we want to show that this is the unique optimal tranport plan. Suppose there were another plan ∇f, where f is a convex function such that the pushforward $\nabla f \# \mu = \nu$ holds. We want to prove they're the same up to a $d\mu$-negligible set. We similarly have that (f, f^*) is an optimal pair for the dual problem. Therefore

$$\int f d\mu + \int f^* d\nu = \int \varphi d\mu + \int \varphi^* d\nu$$

Let π be associated with (φ, φ^*). Then we have

$$\int [f(x) + f^*(y)] d\pi(x, y) = \int [\varphi(x) + \varphi^*(y)] d\pi(x, y) = \int \langle x, y \rangle d\pi(x, y)$$

Since $\pi = (Id \times \nabla\varphi)\#\mu$ (e.g. $T(x) = \nabla\varphi(x)$), we can rewrite

$$\int f(x) + f^*(\nabla\varphi(x)) d\mu(x) = \int \langle x, \nabla\varphi(x) \rangle d\mu(x)$$

and as before, we can subtract and conclude that since the integrand is non-negative by characterization of convex conjugates, it must vanish $d\mu$-almost everywhere. Thus again by

characterization of subdifferential, $\nabla\varphi(x) \in \partial f(x)$ for $d\mu$-almost every x. Since we also had that f is differentiable $d\mu$-almost everywhere, we get $\nabla\varphi(x) = \nabla f(x)$ for $d\mu$-almost evyer x. This completes the proof of the main claim, and in addition to showing a uniqueness of the solution to the Monge-Kantorovich problem, it shows the uniqueness of a gradient $\nabla\varphi$ such that $\nabla\varphi\#\mu = \nu$.

Finally, we quickly want to show that $\nabla\varphi^*$ is an inverse and gives the solution to the Monge-Kantorovich problem of transporting ν onto μ. We have that π is optimal and thus $y = \nabla\varphi(x)$ $d\pi(x,y)$-almost everywhere. Since φ^* is finite $d\nu$-almost everywhere, it is also differentiable $d\nu$-almost everywhere, giving $x = \nabla\varphi^*(y) = \nabla\varphi^*(\nabla\varphi(x))$. This holds $d\mu$ almost everywhere after taking the marginal, and the other way around works the same way.

2.5 Cyclical Monotonicity

We can show another proof via a different approach – cyclical monotonicity. First we give some intuition and motivate the definition of cyclical monotonicity in the discrete case.

2.5.1 Discrete case

The discrete Kantorovich problem can be written as: Minimize

$$\frac{1}{n}\sum_{i,j=1}^{n} \pi_{ij} c(x_i, y_j)$$

where π is a doubly stochastic matrix.

The Monge version of the problem is to simply consider permutations of the points (e.g., find an optimal matching): Minimize

$$\frac{1}{n}\sum_{i=1}^{n} c(x_i, y_{\sigma(i)})$$

where σ is a permutation.

Theorem 2.7. *Choquet's theorem says that this linear minimization problem over bounded convex set has solutions which are the extremal points of the set of doubly stochastic matrices.*

Theorem 2.8. *Birkhoff's theorem says that the extremal points of the set of doubly stochastic $n \times n$ matrices are permutuation matrices.*

Thus the solutions of the Monge and Kantorovich problem coincide in the discrete case.

Therefore, we have that if transport plan $\pi = (1/n)\sum_{i=1}^{n} \delta_{(x_i, y_i)}$ satisfies for all permutations σ

$$\sum_{i=1}^{n} \|x_i - y_i\|_2^2 \leq \sum_{i=1}^{n} |x_i - y_{\sigma(i)}|^2$$

then π is optimal since the optimal transport plan is a permutation and performing no worse than anything in the set of possible optimal plans implies optimality. If π is optimal, then by definition of transport cost it must satisfy this inequality as well, since $y_{\sigma(i)}$ correspond to possible transport plans.

We can see this directly as well:

Theorem 2.9. *The Krein-Milman theorem says that each point of a convex compact set of a Banach space can be written as an average of extremal points of the set.*

Thus, any transference plan can be written as an average of permutation transport plans (e.g., the basis for doubly stochastic matrices is permutation matrices, and the coefficients must be positive and sum to 1). Thus, for arbitrary transport map π_{ij}

$$\frac{1}{n}\sum_{i,j=1}^{n}\pi_{ij}|x_i - y_j|^2 = \frac{1}{n}\sum_{i}^{n}\sum_{k=1}^{m}\alpha_k|x_i - y_{\sigma_k(i)}|^2 \geq \frac{1}{n}\sum_{i}^{n}|x_i - y_{\sigma_{\min}(i)}|^2$$

where $\sigma_{\min}$ is the permutation with the minimizing cost (minimum is less than average) over the m total permutations. Therefore, π_{ij} is only optimal if the cost due to π_{ij} is at most $\frac{1}{n}\sum_i^n |x_i - y_{\sigma_{\min}(i)}|^2$ (this is realizable for π_{ij} simply by choosing the appropriate permutation). This happens iff the cost of π_{ij} is less than the cost for every single permutation.

Note that for $\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n}\delta_{(x_i,y_i)}$, we have that $\pi_{i,j} = 0$ when $j \neq i$, and 1 otherwise, and this $\hat{\pi}$ is an optimal transport plan iff the aforementioned condition holds.

Now, we show this inequality is true iff for all $m \leq N$ and for all $i_1, \cdots, i_m = i_0 \in [N]$,

$$\sum_{k=1}^{m}\|x_{i_k} - y_{i_k}\|_2^2 \leq \sum_{k=1}^{m}\|x_{i_k} - y_{i_{k-1}}\|_2^2$$

First assume this requirement holds. Then we have

$$\sum_{i}\sum_{k=1}^{m}\|x_{i_k} - y_{i_k}\|_2^2 \leq \sum_{i}\sum_{k=1}^{m}\|x_{i_k} - y_{i_{k-1}}\|_2^2$$

since we can choose to split up the n indices into disjoint cycles after fixing a permutation. So this equation holds for any partitioning into permutations. Thus, by definition

$$\sum_{i=1}^{n}\|x_i - y_i\|_2^2 \leq \sum_{i}\sum_{k=1}^{m}\|x_{i_k} - y_{i_{k-1}}\|_2^2$$

Now since σ corresponds to a certain partition into cycles, we have that $\sigma(i_k) = i_{k-1}$ for each cycle in the original permutation, and we get the desired result since the requirement is true for ANY selection of $m \leq N$ and for all $i_1, \cdots, i_k \in [N]$, and thus our arguments hold for any selection of σ.

To get the other direction, we know that for every σ, a decomposition into disjoint cycles is induced. If we consider all possible σ, we will end up seeing that the original equality holds

for every sum over disjoint cycles corresponding to a permutation. All decompositions of all σ will generate every possible cycle (e.g., permutations are generated by decompositions into disjoint cycles). Therefore, to show

$$\sum_{k=1}^{m} \|x_{i_k} - y_{i_k}\|_2^2 \leq \sum_{k=1}^{m} \|x_{i_k} - y_{i_{k-1}}\|_2^2$$

for a specific cycle C, simply consider the permutation which is generated by C and the rest of the cycles are just the identity permutation. Then, write down the original equality:

$$\sum_{i=1}^{n} \|x_i - y_i\|_2^2 \leq \sum_{j=1}^{n-m} \|x_j - y_j\|_2^2 + \sum_{k=1}^{m} \|x_{i_k} - y_{i_k}\|_2^2$$

and canceling both sides gives the result. We can repeat this argument for any specific selection of cycle.

Finally, we want to show that the above inequality for any fixed cycle is equivalent to

$$\sum_{k=1}^{m} \langle y_{i_k}, (x_{i_{k+1}} - x_{i_k}) \rangle \leq 0$$

where $i_{m+1} = i_1$.

$\|y_{i_k}\|_2^2 + \|x_{i_{k+1}} - x_{i_k}\|_2^2 + 2\langle y_{i_k}, x_{i_{k+1}} - x_{i_k} \rangle = \|y_{i_k} - x_{i_k} + x_{i_{k+1}}\|_2^2 = \|x_{i_k} - y_{i_k}\|_2^2 + \|x_{i_{k+1}}\|_2^2 + 2\langle y_{i_k} - x_{i_k}, x_{i_{k+1}} \rangle$

$2\langle y_{i_k}, x_{i_{k+1}} - x_{i_k} \rangle \leq \|x_{i_k} - y_{i_k}\|_2^2 + \|x_{i_{k+1}}\|_2^2 - (\|y_{i_k}\|_2^2 + \|x_{i_{k+1}} - x_{i_k}\|_2^2) + 2\langle y_{i_k} - x_{i_k}, x_{i_{k+1}} \rangle$

$\sum_{k=1}^{m} 2\langle y_{i_k}, x_{i_{k+1}} - x_{i_k} \leq \sum_{k=1} \|x_{i_k} - y_{i_{k-1}}\|_2^2 + \sum_{k=1}^{m} \|x_{i_{k+1}}\|_2^2 - \|y_{i_k}\|_2^2 - \sum_{k=1}^{m} \|x_{i_{k+1}} - x_{i_k}\|_2^2 + \sum_{k=1}^{m} 2(\langle x_{i_{k+1}} y_{i_k} \rangle - \langle x_{i_k}, x_{i_{k+1}}$

where we used the cycle inequality. Now we can take advantage of cyclical symmetry to write

$$= \sum_{k=1}^{m} \|x_{i_k}\|_2^2 - \|y_{i_{k-1}}\|_2^2 + 2\langle x_{i_k}, y_{i_{k-1}} \rangle + \|x_{i_k} - y_{i_{k-1}}\|_2^2 - \sum_{k=1}^{m} \|x_{i_{k+1}} - x_{i_k}\|_2^2 + 2\langle x_{i_{k+1}}, x_{i_k} \rangle$$

Then we expand the squares and cancel things out:

$$= 2\sum_{k=1}^{m} \|x_{i_k}\|_2^2 - \sum_{k=1}^{m} \|x_{i_{k+1}}\|_2^2 + \|x_{i_k}\|_2^2$$

and we take advantage of cyclical symmetry one last time to cancel everything to 0. Thus we have

$$2\sum_{k=1}^{m} \langle y_{i_k}, x_{i_{k+1}} - x_{i_k} \rangle \leq 0$$

as desired. To get the other direction, just reverse the steps.

Note that we have basically defined cyclical monotonicity.

Definition 2.10. Cyclical monotonicity.
A subset $\Gamma \in \mathbb{R}^n \times \mathbb{R}^n$ is cyclically monotone if for all $m \geq 1$, for all $(x_1, y_1), \cdots, (x_m, y_m) \in \Gamma$,

$$\sum_{i=1}^{m} \|x_i - y_i\|_2^2 \leq \sum_{i=1}^{m} \|x_i - y_{i-1}\|_2^2$$

where we take $y_0 = y_m$. Equivalently this can be written as

$$\sum_{i=1}^{m} \langle y_i, x_{i+1} - x_i \rangle \leq 0$$

Now we can see why the definition of cyclical monotonicity is useful. The following partial characterization holds:

Theorem 2.11. *Optimal plans have cyclically monotone support. If the cost is quadratic and π is optimal coupling in the Kantorovich problem, then the support of π is cyclically monotone.*

Proof. (sketch) Let $(x_1, y_1), \cdots, (x_m, y_m)$ be m points in the support of π. We proceed by contradiction and suppose it's not cyclically monotone. Assume

$$\sum \|x_i - y_i\|_2^2 > \sum \|x_i - y_{i-1}\|_2^2$$

Consider balls $B_i(x_i, y_i)$ of mass ϵ under π. Then redefine π by shifting each B_i to a new position (x_i, y_{i-1}) and call the new measure $\hat{\pi}$. X-marginal of $\hat{\pi}$ is still exactly μ, and Y-marginal is approximately ν since we cyclically moved around ϵ masses (e.g., some compensation has happened along the Y-axis). However, the total cost of $\hat{\pi}$ is strictly less than that of π, since by assumption,

$$\sum \|x_i - y_i\|_2^2 - \sum \|x_i - y_{i-1}\|_2^2 > 0$$

Multiply this by ϵ to see what the approximate difference is (since we moved around masses of size ϵ). Thus π wasn't optimal and we have a contradiction. □

We will briefly state the other connections of cyclical monotonicity:

Theorem 2.12. *Rockafellar's theorem.*
A nonempty subset Γ is cycically monotone iff it is included in the subdifferential of a proper lower semi-continuous convex function φ. Morever, maximal cycically monotone subsets (with respect to set inclusion) are exactly the subdifferentials of lower semi-continuous convex functions.

Proof. First we show that the subdifferential set of convex function φ is a cyclically monotone subset of $\mathbb{R}^n \times \mathbb{R}^n$.

Let $(x_1, y_1), \cdots, (x_m, y_m)$ be such that $y_i \in \partial\varphi(x_i)$ for all i. Then by definition of subdifferential,

$$\varphi(z) \geq \varphi(x_i) + \langle y_i, z - x_i \rangle$$

Choosing $z = x_2$ with $i = 1$, $z = x_3$ with $i = 2$, etc. up to $z = x_1$, $i = m$, we get a list of inequalities. Adding them up, we get

$$\sum_{i=1}^{m} \langle y_i, x_{i+1} - x_i \rangle \leq 0$$

which is an equivalent definition of cyclical monotonicity.

For the other direction, check the book. □

From this theorem we can immediately see the connection to Brenier's theorem. This immediately gives us that optimal transport plans are supported on sub-differentials. Note this theorem holds even for non quadratic losses! This is true on an arbitrary Hilbert space, we didn't use properties of quadratic loss anywhere.

If we now take advantage of properties of differentiability on convex sets (as before), we get that there is an optimal transport map that's a gradient of a convex function, as before (but this does not yet prove uniqueness).

2.6 Uniqueness of Gradient Map

To prove uniqueness, one can use Aleksandrov's lemma.

Lemma 2.13. *Aleksandrov's lemma.*
Let $\varphi, \bar{\varphi}$ be convex functions s.t. $\varphi(x_0) = \bar{\varphi}(x_0)$, but $\nabla\varphi(x_0) \neq \nabla\bar{\varphi}(x_0)$. Let $V = \{\varphi > \bar{\varphi}\}$ and

$$Z = \nabla\bar{\varphi}^{-1}(\nabla\varphi(V))$$

Then, $x_0 \in \bar{V}, Z \subset V$, but Z lies a positive distance away from x_0: $\mu(Z) < \mu(V)$.

With this lemma, it is possible to complete the proof of Brenier's theorem without any assumptions of finite second moments, as we required before in the duality argument.

The idea is again you assume for sake of contradiction that $\nabla\varphi$ and $\nabla\bar{\varphi}$ are not equal on the support of μ while having the same pushforward $\nabla\varphi\#\mu = \nabla\bar{\varphi}\#\mu = \nu$. Then let x_0 be in the support of μ and assume that $\varphi(x_0) = \bar{\varphi}(x_0)$. Using a nonsmooth implicit function theorem for convex functions, one can show that the measure of $\{\varphi = \bar{\varphi}\}$ is small under μ. We chose x_0 in the support of μ, so it's still possible to find some small neighborhood intersecting (WLOG) $V = \{\varphi > \bar{\varphi}\}$. This sets up the application of Aleksandrov's lemma to deduce $\nabla\varphi$ and $\nabla\bar{\varphi}$ don't have the same pushforward, a contradiction. In particular,

$$\nabla\bar{\varphi}\#\mu[\nabla\varphi(V)] = \mu[\nabla\bar{\varphi}^{-1}(\nabla\varphi(V))] = \mu[Z] < \mu[V] \leq \mu[\nabla\varphi^{-1}(\nabla\varphi(V))] = \nabla\varphi\#\mu[\nabla\varphi(V)]$$

where we used Alexandrov's lemma in the middle.

3 Beyond Quadratic Costs

Let c be some cost function. The key idea here is the generalization of duality to the notions of c-concavity, c-superdifferential, and the c-transform.

Definition 3.1. c-concavity. A function φ is c-concave if there exists ψ such that for all x,

$$\varphi(x) = \inf_y [c(x,y) - \psi(y)]$$

Definition 3.2. c-cyclically monotone. For any $(x_1, y_1), \cdots, (x_m, y_m)$,

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

with convention $y_0 = y_m$.

Definition 3.3. c-superdifferential. $\partial^c \varphi$ of c-concave function φ is defined as the set

$$\partial^c \varphi(x) = \{y : \forall z, \varphi(z) \leq \varphi(x) + [c(z,y) - c(x,y)]\}$$

Definition 3.4. c-transform.

$$\varphi^c(y) = \inf_x [c(x,y) - \varphi(x)]$$

Definition 3.5. Duality.
We have that

$$\varphi(x) + \varphi^c(y) \leq c(x,y)$$

Definition 3.6. Generalized Rockafellar (Ruschendorf's theorem).
Any c-cyclically monotone set can be included in the c-superdifferential of a c-concave function.

3.1 Strictly Convex Case

A strictly convex cost is defined $c(x,y) = c(x-y)$ where c is strictly convex and superlinear on $\mathbb{R}^n$. Then, when ∇c is invertible with Legendre transform ∇c^*, then if φ is c-concave and differentiable at x,

$$\partial^c \varphi(x) = \{x - \nabla c^*(\nabla \varphi(x))\}$$

Note also that ∇c^* is an inverse of ∇c.

Now,, there exists a unique optimal transport plan which is uniquely determined in measure μ and satisfies $T \# \mu = \nu$ with

$$T(x) = x - \nabla c^*(\nabla \varphi(x))$$

The idea for the proof, as you may have gathered from the definitions presented previously, is to show that

$$\partial^c \varphi(x) = \{x - \nabla c^*(\nabla \varphi(x))\}$$

is a singleton set. Morally you will follow a similar outline of proof.

3.2 Strictly Concave Case

If $c(x, y) = c(|x - y|)$ and c is strictly concave, then a similar theorem holds, with the exception that any optimal transport plan must require the mass which is shared between μ and ν to stay in the same place. After ensuring this, can apply transport map

$$T(x) = x - (\nabla c)^{-1}(\nabla \varphi(x))$$

to the rest of the mass.

3.3 Other generalizations

It's also possible to extend to Riemannian manifolds.