K-means clustering, and basics of Optimal Transport

MATH 697 AM:ST

November 28th, 2017

Warmup

Let $C = \{x_1, \ldots, x_N\}$ be a subset of $\mathbb{R}^p$.

The center of mass of C, is defined as

$$\bar{x}_C := \frac{1}{N} \sum_{i=1}^{N} x_i$$

As is known, $\bar{x}_C$ has a variationa characterization

$$\bar{x}_C = \operatorname*{argmin}_{x \in \mathbb{R}^p} \sum_{i=1}^{N} \|x_i - x\|^2$$

Warmup

Another useful property of $\bar{x}_C$ is the following: suppose C is such that $\bar{x}_C = 0$, and let us consider the sum

$$\sum_{i,j=1}^{N} \|x_i - x_j\|^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} \|x_i - x_j\|^2$$

Then, note that for each i we have

$$\begin{aligned} \sum_{j=1}^{N} \|x_i - x_j\|^2 &= \sum_{j=1}^{N} \|x_i\|^2 - 2(x_i, x_j) + \|x_j\|^2 \\ &= N\|x_i\|^2 + \sum_{j=1}^{N} \|x_j\|^2 \end{aligned}$$

Warmup

Adding this up in i, we have

$$\sum_{i,j=1}^{N} \|x_i - x_j\|^2 = N \sum_{i=1}^{N} \|x_i\|^2 + N \sum_{j=1}^{N} \|x_j\|^2$$
$$= 2N \sum_{i=1}^{N} \|x_i\|^2$$

What if $\bar{x}_C \neq 0$? Then we can substract $\bar{x}_C$ from each $x_i \ldots$

Warmup

We conclude that $\bar{x}_C$ has the following property for any set C:

$$\frac{1}{2N}\sum_{i,j=1}^{N}\|x_i - x_j\|^2 = \sum_{i=1}^{N}\|x_i - \bar{x}_C\|^2$$

In other words: if one wants, for whatever reason, to add up the pairwise square distances between points in a finite set, then up to a factor of $2N$ this is the same as adding up the square distances from a point in C to its center of mass.

Clustering
(RECAP)

In clustering, one is given **disorganized** data

Clustering

Examples

(RECAP)

. . . and then one seeks to organize it in terms of groups of similar elements

Clustering

Examples

(RECAP)

Ultimately, this problem comes down to dividing discrete sets in a proper metric space

Key point: No training labels were given a priori!

Clustering

Challenges

(RECAP)

1. Typically, these are **unsupervised** problems: no labels are given in advance.
2. In fact, unclear a priori what is the "right" number of clusters and what their nature is (e.g. I give you a data set of all the music available on some streaming service, can one automatically generate playlists with similar songs?)
3. Clustering in itself is a form of dimensional reduction for a data set by way of "coarsening".
4. Often used in combination with other learning algorithms.

Clustering

Similarity measure / Proximity matrices

(RECAP)

One then has a set $\{x_1, \ldots, x_N\}$, where one is given quantitative data measuring how dissimilar two elements are

$$w(x_i, x_j) = \begin{cases} \text{the greater this number,} \\ \text{the greater the similarity} \\ \text{between } x_i \text{ and } x_j \end{cases}$$

Clearly, this amounts to a weighted graph!!

From Harmonic Functions to Spectral Clustering
(RECAP)

For a weighted (G, w_{ij}), we want to partition G in two classes

$$G = A \cup B, \quad A \cap B = \emptyset$$

We want to make this partition as optimal as possibe: one criterium, the "connectivity" between the two components of the partition must be minimal –but in what sense?

Given two sets $A, B \subset G$ (disjoint or not) we define their **cut**

$$\text{cut}(A, B) = \sum_{x \in A} \sum_{y \in B} w_{xy}$$

From Harmonic Functions to Spectral Clustering

Normalized Cuts

(RECAP)

Shi and Malik (2000) proposed the following minimization problem, modifying Leahy and Wu's proposal:

$$\mathrm{Ncut}(A, B) = \mathrm{cut}(A, B) \left(\frac{1}{\mathrm{Vol}(A)} + \frac{1}{\mathrm{Vol}(B)} \right)$$

This gives a criterium to divide a data set G into **two** distinguished sets, which avoids the tendency to pick small, disconnected subsets.

From Harmonic Functions to Spectral Clustering

Normalized Cuts

(RECAP)

The hope is that the minimizing partition (A_0, B_0) captures two important features of the data set. This is a special instance of a **clustering problem**.

$$\mathrm{Ncut}(A, B) = \mathrm{cut}(A, B)\left(\frac{1}{\mathrm{Vol}(A)} + \frac{1}{\mathrm{Vol}(B)}\right)$$

This minimization problem, it must be noted, is **NP-hard**.

If one is willing to give up a small amount of accuracy, it can be approximated via a problem which is solvable in polynomial time (or, as fast as one compute find graph eigenfunctions).

From Harmonic Functions to Spectral Clustering

Ncut and indicator functions

(RECAP)

Take $A \subset G$, and consider the function

$$f(x) = \chi_A(x)$$

Then, observe that

$$\begin{aligned} \text{cut}(A, A^c) &= \sum_{x \in A, y \in A^c} w_{xy} \\ &= \sum_{x,y \in G} w_{xy} f(x)(1 - f(y)) \\ &= \frac{1}{2} \sum_{x,y \in G} w_{xy} (f(x) - f(y))^2 \end{aligned}$$

From Harmonic Functions to Spectral Clustering

A relaxation of the graph cut problem

(RECAP)

Shi and Malik observed that if we define

$$u = f - b(1 - f), \quad \text{where } b = \frac{\text{Vol}(A)}{\text{Vol}(G)}$$

Then, A minimizes $\text{Ncut}(A, A^c)$ if and only if u minimizes

$$\frac{\langle \Delta u, u \rangle}{\langle u, u \rangle_d}$$

over the set of u's satisfying the constraints

$$u(x) \in \{1, -b\} \; \forall \; x, \;\; \langle u, \mathbf{1} \rangle_d = 0.$$

From Harmonic Functions to Spectral Clustering

A relaxation of the graph cut problem

(RECAP)

Now, this last problem is precisely equivalent to the Ncut minimization problem.

The **relaxation** consists in looking for a minimizer of

$$\frac{\langle \Delta u, u \rangle}{\langle u, u \rangle_d}$$

over all functions $u : G \mapsto \mathbb{R}$, **ignoring the constraints.**

This relaxed problem is the same as finding the first non-zero eigenfunction of the associated Laplacian. This problem can be solved in polynomial time.

From Harmonic Functions to Spectral Clustering

A relaxation of the graph cut problem

Of course, there is a drawback: the relaxed problem is of course not the same as the original one.

Having solved the relaxed problem, however, we can approximate the original solution to a great degree.

Simply, one takes as a guess

$$A = \{x \in G \mid \phi_2(x) > 0\}$$

(one can also take instead $\{\phi_2(x) > \delta\}$ for some other value δ)

Shi-Malik Ncut minimization

A relaxation of the graph cut problem

Shi-Malik Ncut minimization

A relaxation of the graph cut problem

Shi-Malik Ncut minimization

A relaxation of the graph cut problem

Clustering

Dissimilarity matrices

Often, instead of thinking of proximity/similarity between nodes, it is more convenient to think in terms of how dissimilar they are. That is, one is given a **dissimilarity matrix**

$$d_{ij} = \text{ the larger this number is, the more dissimilar } i \text{ and } j$$

The d_{ij} should be thought of as a distance, while the weight matrix w_{ij} in a graph is an inverse power of d_{ij}: if the w_{ij} is larger, that means that i and j have a stronger connection or similarity between them. Accordingly, in given a weighted graph it is natural to take w_{ij}^{-1} as a dissimilarity matrix.

Clustering

K-means

This algorithm is extremely popular, and it arises when

$$G = \{x_1, \ldots, x_N\} \subset \mathbb{R}^p$$

and the dissimilarity matrix is given by the Euclidean distance squared,

$$d(x_i, x_j) = d_{ij} = \|x_i - x_j\|^2$$

Clustering

K-means

Let C be a partition of G with K elements, that is

$$C = C_1 \cup \ldots \cup C_K$$

with the C_k being pairwise disjoint.

Then, the loss function takes the form,

$$J(C) = \sum_{k=1}^{K} \sum_{x,y \in C_k} \|x - y\|^2$$

Clustering

K-means

As we saw in the discussion at the beginning of today's class,

$$\begin{aligned}\sum_{x,y\in C_k} \|x-y\|^2 &= 2N \sum_{x\in C_k} \|x-\bar{x}_{C_k}\|^2 \\ &= 2N \min_{m\in\mathbb{R}^p} \sum_{x\in C_k} \|x-m\|^2\end{aligned}$$

Clustering

K-means

In this manner we arrive at the following minimization problem: find a partition C with K elements, as well as K points $m_1, \ldots, m_K \in \mathbb{R}^p$, minimizing the expression

$$\sum_{k=1}^{N} \sum_{x \in C_k} \|x - m_k\|^2$$

Thus we arrive to the K-means algorithm.

Clustering

The K-means algorithm

Input: Points $x_1, \ldots, x_N$ and a number $K < N$.

1. Initialize a partition $C^{(0)}$ with K elements in any way.
2. (Loop) We start with $C^{(n)}$ $(n = 0, 1, 2, \ldots)$
   - 2.1 Compute the means $m_{C_1^{(n)}}, \ldots, m_{C_K^{(n)}}$ for $C^{(n)}$'s clusters.
   - 2.2 Let $C^{(n+1)}$ be the partition given by

$$C_k^{(n+1)} = \{x_i \mid \|x_i - m_{C_k^{(n)}}\| \leq \|x_i - m_{C_j^{(n)}}\| \ \ \forall\, j\}$$

   for $k = 1, \ldots, K$.
3. Stop after some predetermined number of steps or after some threshold is met.

Clustering

K-means: Practical considerations

1. The K-means algorithm may produce different local minima —this problem is far from being a convex one!
2. There are often multiple global minima. To see an example of this, consider a data set with symmetries, e.g. consider $2N$ evenly placed points on the unit circle, multiple solutions to K-means with $K = 2$ clusters!
3. To avoid being trapped by local minima, a common practice is to generate several initial guesses at random, and run the algorithm for each one of them –then pick the cluster with the smallest possible value.

Clustering

K-medoids

There are variations of the K-means algorithm when the dissimilarity matrix is not the squared Euclidean distance. The K-medoids algorithm directly uses the Euclidean distance

$$d(x_i, x_j) = d_{ij} = \|x_i - x_j\|$$

Since there is no square here, one cannot make the same algebraic reduction as in K-means.

Clustering

The K-medoids algorithm

Input: Points $G = \{x_1, \ldots, x_N\}$ and a number $K < N$.

1. Initialize a partition $C^{(0)}$ with K elements in any way.
2. (Loop) We start with $C^{(n)}$ $(n = 0, 1, 2, \ldots)$
   2.1 For each cluster of $C^{(n)}$, let m_k be chosen so that

$$m_k = \text{argmin}_{m \in C_k^{(n)}} \sum_{x_i \in C_k^{(n)}} \|x_i - m\|$$

   2.2 Let $C^{(n+1)}$ be the partition given by

$$C_k^{(n+1)} = \{x_i \mid \|x_i - m_{C_k^{(n)}}\| \leq \|x_i - m_{C_j^{(n)}}\| \ \ \forall\, j\}$$

   for $k = 1, \ldots, K$.
3. Stop after some predetermined number of steps or after some threshold is met.

Image compression as dimensionality reduction

A popular use of clustering is in image compression: one replaces a gray scale image that say, contains 1000 different scales of gray, with one that only has K different scales of gray, for a much smaller K. This is an instance of what is often known as **vector quantization**.

Image compression as dimensionality reduction

Image compression as dimensionality reduction

(Some compression: $K = 200$ scales of gray used)

Image compression as dimensionality reduction

(Higher compression: only $K = 4$ scales of gray used)

Information Geometry

For our next topic, we shall make a quick comparison between various metrics over the set of distributions and probability measures.
First we have, of course, the L^1 distance

$$d_{L^1}(f,g) = \int_{\mathbb{R}^d} |f(x) - g(x)| \, dx$$

and the L^p distance ($p \in [1, \infty]$),

$$d_{L^p}(f,g) = \left(\int_{\mathbb{R}^d} |f(x) - g(x)|^p \, dx \right)^{\frac{1}{p}}$$

Information Geometry

Another important metric, specially in image processing, is the **total variation** (TV) distance/norm

$$d_{TV}(f, g) = \|f - g\|_{TV}$$

Where the TV norm of a function $w \in L^1(\mathbb{R}^d)$ is defined by

$$\sup \left\{ \int_{\mathbb{R}^d} \operatorname{div}(\phi(x)) w(x)\, dx \;\mid\; \|\phi\|_{L^\infty} \leq 1 \right\}$$

Information Geometry

There is also the recently popular **Earth Mover Distance**: assume f and g are both non-negative and have total mass 1, then,

$$d_{EMD}(f,g) = \inf_{\pi} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y| \, d\pi(x,y)$$

The infimum being over all measures π such that

$$\pi(A \times \mathbb{R}^d) = \int_A f(x) \, dx$$
$$\pi(\mathbb{R}^d \times B) = \int_B g(y) \, dy$$

Information Geometry

Last but not least, not exactly a metric, but in many ways simillar: once again assume f and g are non-negative and have total mass 1, then the **relative entropy**, aka the Kullback-Leibler divergence, is defined as

$$\begin{aligned} D_{KL}(f \mid g) &= \int_{\mathbb{R}^d} f(x) \log\left(\frac{f(x)}{g(x)}\right)\, dx \\ &= H(f \mid g) - H(f) \end{aligned}$$

where $H(f \mid g)$ and $H(f)$ are the cross-entropy of f and g, and entropy, respectively, defined by

$$H(f \mid g) := -\int_{\mathbb{R}^d} f(x) \log(g(x))\, dx, \quad H(f) := H(f \mid f)$$

Basics of Optimal Transport

MATH 697 AM:ST

November 28th, 2017

Warmup

or: a meta-mathematical discussion

Warmup

or: a meta-mathematical discussion

Warmup

or: a meta-mathematical discussion

Warmup

or: a meta-mathematical discussion

Warmup

or: a meta-mathematical discussion

A way to judge various arrangements (define "what is best?")

$$
\begin{aligned}
N &= \#\text{ of blocks in the city} \\
c_{ij} &= \text{ distance from block } i \text{ to block } j \\
f_i &= \text{ population in block } i \\
\sigma(i) &= \text{ block \# of fire dept assigned to block } i
\end{aligned}
$$

One possible criteria:

"Best" := $\sigma(\cdot)$ minimizes the average distance

$$\sum_{k=1}^{N} c_{i\sigma(i)} f_i$$

Warmup

or: a meta-mathematical discussion

A way to judge various arrangements (define "what is best?")

$$\begin{aligned} N &= \#\text{ of blocks in the city} \\ c_{ij} &= \text{ distance from block } i \text{ to block } j \\ f_i &= \text{ population in block } i \\ \sigma(i) &= \text{ block \# of fire dept assigned to block } i \end{aligned}$$

One possible criteria:

"Best" := $\sigma(\cdot)$ minimizes the average distance

$$\frac{1}{M}\sum_{k=1}^{N} c_{i\sigma(i)} f_i \quad \text{where} \quad M := \sum_{i=1}^{N} f_i.$$

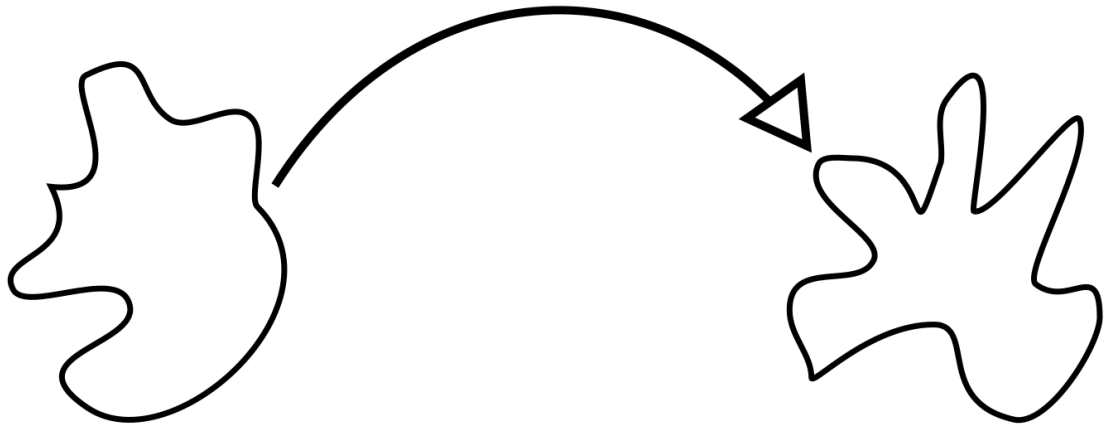
(ok fine, **this** is the actual average)

Optimal Transport as a field of mathematics

(it's not *just* about about actual transport!)

"Optimal transport" is a field of mathematics & an array of tools increasingly used throughout engineering and science.

It's concerned with maps that preserve volume and/or ways of "matching" two distributions together



and this "transport"/"matching" is optimal *in some sense*.

Optimal Transport as a field of mathematics

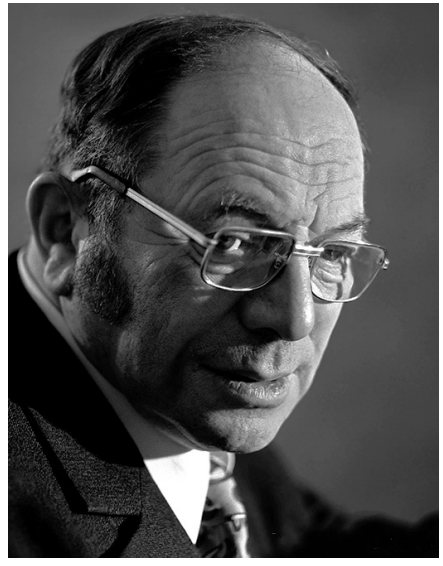
Goes back to the 19th century, mostly dormant during 20th..

1. Gaspard Monge (19th century, French physicist and mathematician)
2. Leonid Kantorovich (20th century, only Soviet Nobel prize in economics!)

The field has really blossomed since the early 90's and has become a fairly developed field whose methods are used in:

image processing, pattern recognition, fluid mechanics, weather modelling, geometric optics, probability, statistical inference, differential geometry, economics...

Optimal Transport as a field of mathematics



Gaspard Monge (1746-1818) & Leonid Kantorovich (1912-1986)

The Monge Problem

Setup

We are given two sets of N points in $\mathbb{R}^d$

$$X = \{x_1, \ldots, x_N\}, \quad Y = \{y_1, \ldots, y_N\}$$

and consider the set of bijections

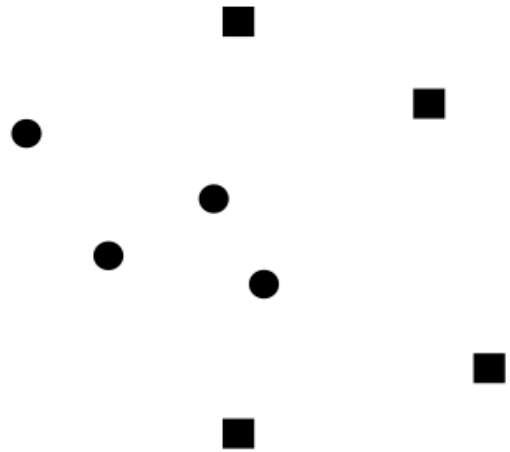
$$T : X \to Y$$

Think of X as denoting the location of units of some material, and Y location where these units want to be transported. A map T then represents a transportation map that tells you to which y_j to transport the unit located at a given x_i.

The Monge Problem

Setup

Let's denote points in X by and those in Y by circles.



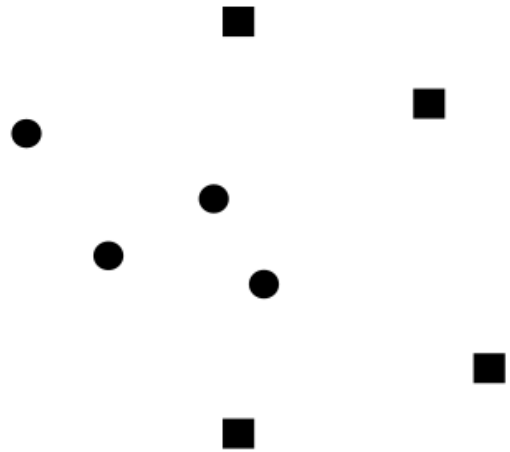
There could be a cost associated to transporting the material, proportional to distance,

$$|x_i - T(x_i)|$$

The Monge Problem

Setup

Let's denote points in X by and those in Y by circles.



In general, we have a function

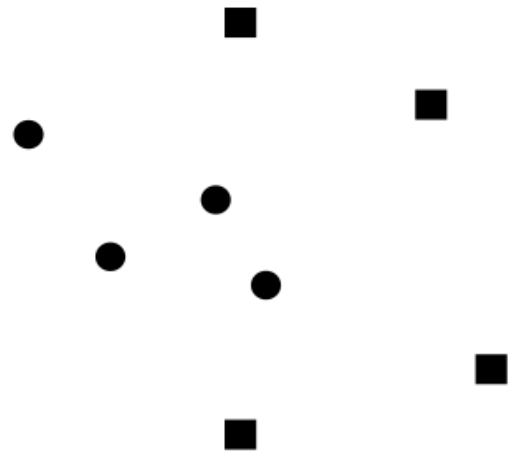
$$c : X \times Y \mapsto \mathbb{R}$$

called the **transportation cost** from x to y, denoted $c(x, y)$.

The Monge Problem

Setup

Let's denote points in X by and those in Y by circles.



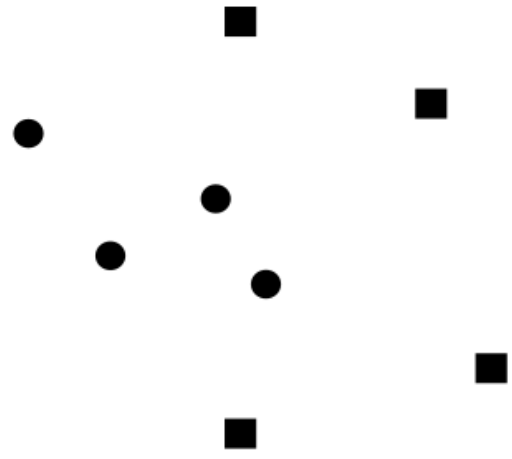
Tthen we may want to minimize the **total transportation cost**, given by the functional

$$J(T) = \sum_{i=1}^{N} c(x_i, T(x_i))$$

The Monge Problem

Setup

Problem



Find a bijective function (a "map") $T : X \mapsto Y$ minimizing

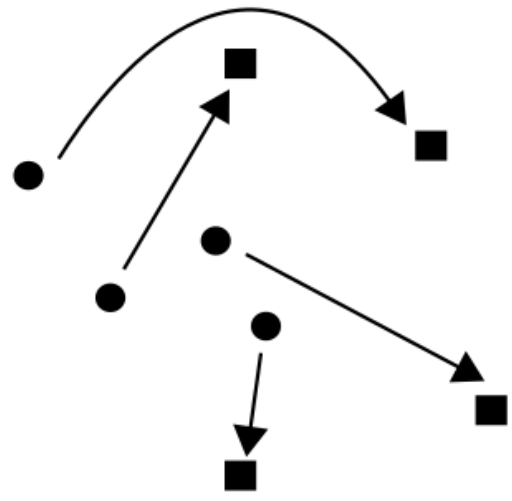
$$J(T) = \sum_{x \in X} c(x, T(x))$$

$$(\textit{Total transportation cost of } T)$$

The Monge Problem

Setup

Problem



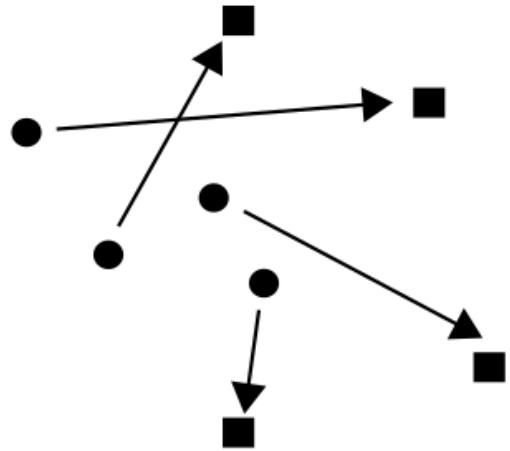
Find a bijective function (a "map") $T : X \mapsto Y$ minimizing

$$J(T) = \sum_{x \in X} c(x, T(x))$$

(*Total transportation cost of* T)

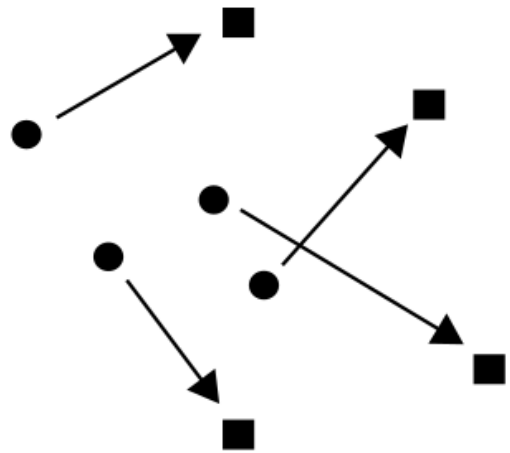
The Monge Problem

As X and Y are finite, there are finitely many choices, so at least one of them achieves the smallest value



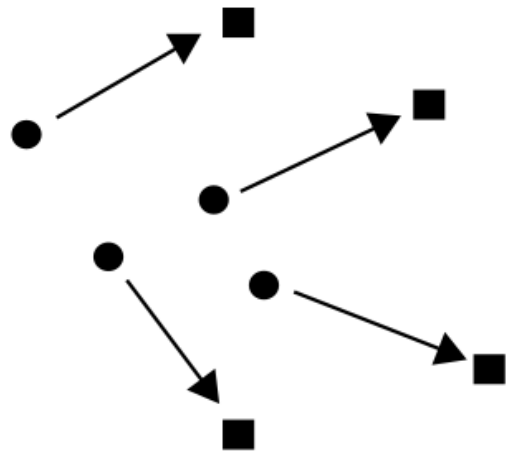
The Monge Problem

As X and Y are finite, there are finitely many choices, so at least one of them achieves the smallest value



The Monge Problem

As X and Y are finite, there are finitely many choices, so at least one of them achieves the smallest value



The Monge Problem

Characterizing optimality

Let T be an optimal transport map. Then, if $n > 0$, $x_1, \ldots, x_n$ are points in X, and σ is a permutation of $\{1, \ldots, n\}$,

$$\sum_{k=1}^{N} c(x_k, y_k) \leq \sum_{k=1}^{N} c(x_k, T(x_{\sigma(k)}))$$

The Monge Problem

Characterizing optimality

Let T be an optimal transport map. Then, if $n > 0$, $x_1, \ldots, x_n$ are points in X, and σ is a permutation of $\{1, \ldots, n\}$,

$$\sum_{k=1}^{N} c(x_k, y_k) \leq \sum_{k=1}^{N} c(x_k, T(x_{\sigma(k)}))$$

SPOILER ALERT: This is a very strong property! A theorem of Rockafellar says (morally) that if $c(x, y) = |x - y|^2$,

$$T(x) = \nabla u(x)$$

where u is a convex scalar function (more on this later).

The Monge Problem

There is a $N \to \infty$ limit to this problem: instead of two finite sets X and Y we are given two mass distributions $f, g : \mathbb{R}^d \to \mathbb{R}$, with equal total mass, and we seek to minimize the functional

$$J(T) = \int_{\mathbb{R}^d} c(x, T(x)) f(x)\, dx$$

over the set of T's which are **measure preserving**, namely

$$\int_{T(A)} g(y)\, dy = \int_A f(x)\, dx \;\forall\; A.$$

The Monge Problem

Setup –continuous case

We are given

1. Two distributions $f(x)$ and $g(y)$ with

$$\int_{\mathbb{R}^d} f(x)\,dx = \int_{\mathbb{R}^d} g(y)\,dy$$

A transformation $T : X \mapsto Y$ will be called a **transport map** if

$$\int_E f(x)\,dx = \int_{T(E)} g(y)\,dy$$

for all Borel sets $E \subset X$

The Monge Problem

Setup –continuous case

We are given

2. A cost function $c(x, y) : X \times Y \mapsto \mathbb{R}$ representing the cost of transporting one unit of mass from location x to location y.

For a transport map T, the **total transportation cost** is

$$J(T) := \int_X c(x, T(x)) f(x)\, dx$$

Monge's Problem:

Among all transport maps from f to g, find one for which the total transportation cost is minimized.

The Kantorovich Problem

Setup

A variation

We are now given "mass densities" $f, g : X, Y \mapsto \mathbb{R}$, representing

$$\begin{aligned} f(x) &= \text{ output at } x \\ g(y) &= \text{ demand at } y \end{aligned}$$

Their total masses are equal

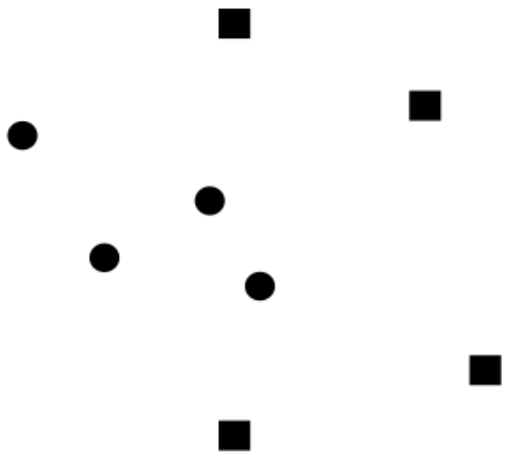
$$\sum_{x \in X} f(x) = \sum_{y \in Y} g(y)$$

The Kantorovich Problem

Setup

Looking for a function $T : X \mapsto Y$ that "preserves measure" is asking for too much here! There may be no such functions T!.

Solution: allow ourselves **to split the mass** leaving each x

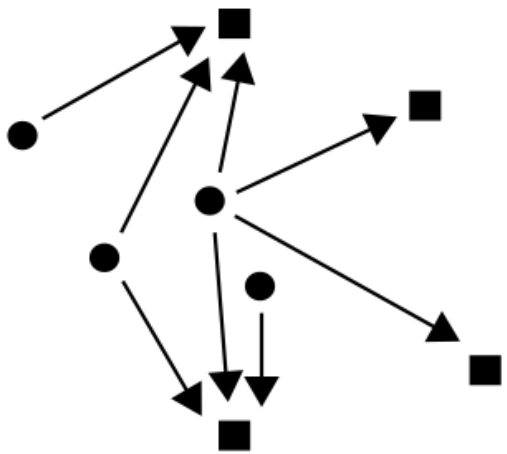


The Kantorovich Problem

Setup

Looking for a function $T : X \mapsto Y$ that "preserves measure" is asking for too much here! There may be no such functions T!.

Solution: allow ourselves **to split the mass** leaving each x



The Kantorovich Problem

Setup

This splitting we encode by a function $\pi(x, y)$, interpreted as

$$\pi(x, y) = \text{ mass sent from } x \text{ to } y$$

$$f(x) = \sum_{y \in Y} \pi(x, y), \quad g(x) = \sum_{x \in X} \pi(x, y)$$

Such a π will be called a **transport plan** between f and g.

The Kantorovich Problem

Setup

Then, we are looking among all transport plans $\pi : X \times Y \mapsto \mathbb{R}$ one which minimizes the total transport cost

$$J(\pi) = \sum_{x \in X} \sum_{y \in Y} c(x, y)\pi(x, y)$$

This is known as the Kantorovich problem.

Note that **this is a linear optimization problem in** π: the set of admissible π's is a convex set in $\mathbb{R}^N$, with $N = |X||Y|$, and the objective functional $J(\pi)$ is linear.

The Monge-Kantorovich Problem

As it turns out, this formulation of the problem, due to Kantorovich, is not only interesting in its own right, it is also the best approach towards solving the Monge problem.

The Monge-Kantorovich Problem

Once again, but more generally

1. Compact domains X and Y, with distributions (i.e. Borel probability measures) μ and ν on each of them.
2. A continuous function $c : X \times Y \mapsto \mathbb{R}$.

In this more general context, a **transportation plan** between μ and ν is measure π in $X \times Y$ such that

$$\pi(A \times Y) = \mu(A), \;\; \pi(X \times B) = \nu(B).$$

The Monge-Kantorovich Problem

Once again, but more generally

In analogy with the discrete setting, one defines the **total transportation cost** for a plan π

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

Kantorovich's Problem: *Among all transport plans between ν and μ, find one for which the total transportation cost is minimized.*

The Earth Mover Distance

There is a popular metric n computer science (since about the last decade), directly related to the Kantorovich problem. This is the **Earth Mover Distance (EMD)**: assume f and g are both non-negative and have total mass 1, then,

$$d_{EMD}(f,g) = \inf_{\pi} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y| \, d\pi(x,y)$$

The infimum being over all measures π such that

$$\pi(A \times \mathbb{R}^d) = \int_A f(x) \, dx,$$
$$\pi(\mathbb{R}^d \times B) = \int_B g(y) \, dy$$

The Earth Mover Distance

There is a popular metric n computer science (since about the last decade), directly related to the Kantorovich problem. This is the **Earth Mover Distance (EMD)**: assume f and g are both non-negative and have total mass 1, then,

$$d_{EMD}(f,g) = \inf_{\pi} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y| \, d\pi(x,y)$$

That is, the EMD between f and g is the optimal value in the respective Kantorovich problem with Monge cost

$$c(x,y) = |x-y|$$

The Monge-Kantorovich Problem

Monge versus Kantorovich

Monge:

- Highly nonlinear variational problem. How do we find a minimizer?
- Solutions are relatively nice: single valued functions!.

Kantorovich:

- Amounts to minimizing a linear functional over a bounded convex set in (it's a linear program!).
- Solutions may be "multi-valued" (mass-splitting!).
- It is a **relaxation** of the Monge problem.

The Monge-Kantorovich Problem

Every transport map determines a transport plan

Further assumptions: $X, Y \subset \mathbb{R}^d$, μ has compact support and it is absolutely continuous with respect to Lebesgue measure.

Theorem (Brenier) *The optimal transport plan is unique and given by a map T. Moreover, $\exists\ u : X \mapsto \mathbb{R}$, convex, such that*

$$T(x) = \nabla u(x) \quad \mu - \text{a.e.}$$

Optimal transport with discrete target measures

Example

Let X, Y be bounded domains in $\mathbb{R}^d$, $\{y_1, \ldots, y_N\} \subset Y$, and

$$\mu = \frac{1}{|X|}\, dx, \quad \nu = \frac{1}{N} \sum_{k=1}^{N} \delta_{y_k}$$

Then, the optimal transport map $T : X \mapsto Y$ is such that

$$T(x) = \nabla u(x) \ \text{ a.e.}$$

where, for some numbers $z_1, \ldots, z_N$

$$u(x) = \max_k \{-x \cdot y_k + z_k\}$$

Optimal transport with discrete target measures

Example

What is the meaning of $u(x) = \max_k \{-x \cdot y_k + z_k\}$??

Well, for once, for every j, we have

$$\begin{aligned} T^{-1}(y_j) &= \{x \in X \mid -x \cdot y_j + z_j \leq -x \cdot y_k + z_k \ \ \forall\, k\} \\ &= \{x \in X \mid 0 \leq x \cdot (y_j - z_k) + z_k - z_j \quad \forall\, k\} \end{aligned}$$

In particular (say, if X is convex) every preimage $T^{-1}(y_j)$ is a convex set. Moreover, if j is such that $T^{-1}(y_j)$ is compactly contained in X, then this set will in fact be a convex polytope.

Optimal transport: the role of convex potentials

Why is Brenier's theorem true? Think of what optimality entails for the discrete case!

Optimal transport: the role of convex potentials

Why is Brenier's theorem true? Think of what optimality entails for the discrete case!

In fact, for the Monge-Kantorovich problem, the following holds:

Lemma: *Let π be an optimal transport plan, and take N points $(x_k, y_k) \in \mathrm{spt}(\pi)$. Then, for any permutation σ we have*

$$\sum_{k=1}^{N} c(x_k, y_k) \leq \sum_{k=1}^{N} c(x_{\sigma(k)}, y_k)$$

Optimal transport and partitioning

Take $\mu \in \mathcal{P}(\Omega)$ where $\Omega \subset \mathbb{R}^d$ and some number N. We are looking for a partition of Ω into N pieces of equal measure:

$$\Omega = \bigcup_{k=1}^{n} S_k, \quad \mu(S_k) = \frac{1}{N} \text{ for } k = 1, \ldots, N$$

The optimality criterion being to minimize

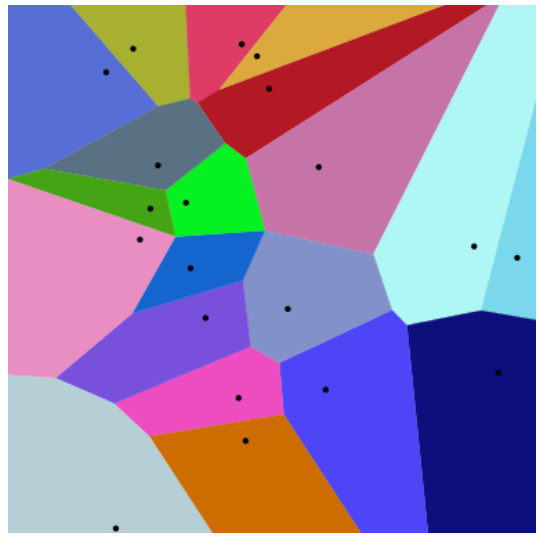
$$\sum_{k=1}^{N} \int_{S_k} c(x, y_k)\ \mu(x)$$

The point is that the N points $\{y_1, \ldots, y_N\}$ are part of the unknowns.

Optimal transport and partitioning

Contrast this with: Voronoi diagrams

Consider a region Ω and points $\{x_1, \ldots, x_N\}$.



These points yield a "partition" of Ω into N regions, known as its Voronoi diagram, its cells being defined by

$$D_k = \{x \in \Omega \ \mid \ |x - x_k| \leq |x - x_j| \ \ j = 1, \ldots, N\}$$