Numerical Optimal Transport and its Applications

Gabriel Peyré
CNRS & DMA
École Normale Supérieure
gabriel.peyre@ens.fr

Abstract

Optimal transport is an old problem, formulated by Monge in the 18th century. However, it took several mathematical revolutions for it to become an indispensable tool both in theory and in practice. This article traces these revolutions, initiated by Leonid Kantorovich during the Second World War. His formulation lends itself to advanced mathematical analysis and its application to many problems. It makes optimal transportation a tool of choice for addressing the recent explosion of data science.

1 Optimal Transport of Monge

Gaspard Monge, in addition to being a great mathematician, took an active part in the French Revolution, and created the École Polytechnique as well as the École Normale Supérieure. Motivated by military applications, he formulated in 1781 the problem of optimal transport [6]. He asked himself the question of how to calculate the most economical way of transporting soil between two places to make embankments. In his original text, he made the assumption that the cost of moving a unit of mass is equal to the distance traveled, but one can use any cost adapted to the problem to be solved.

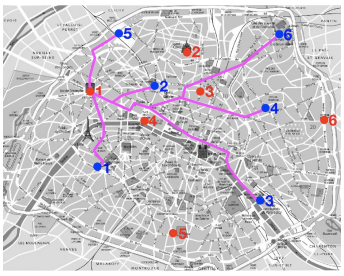
1.1 Monge's problem

To illustrate the problem and its mathematical formulation, let's look at the optimal way of distributing croissants from bakeries to cafés in the morning in Paris. For simplicity, we will assume that there are only six bakeries and cafés, which can be seen in Figure 1 (bakeries are in red and cafés in blue). The cost to be minimized is the total journey time, and we note $C_{i,j}$ the time between the bakery $i \in \{1, \dots, 6\}$ and the café $j \in \{1, \dots, 6\}$. For example, we have $C_{3,4} = 10$, which means that there is a ten-minute commute between bakery number 3 and café number 4.

In order to meet the supply constraint (also known as mass conservation), each bakery must be connected to one and only one café. As there are the same number of bakeries as café, this implies that each café is also connected to one and only one bakery. We will note

$$\sigma : i \in \{1, \dots, 6\} \longmapsto j \in \{1, \dots, 6\}$$

c_{ij}	y_1	y_2	y_3	y_4	y_5	y_6
x_1	12	10	31	27	10	30
x_2	22	7	25	15	11	14
x_3	19	7	19	10	15	15
x_4	10	6	21	19	14	24
x_5	15	23	14	24	31	34
x_6	35	26	16	9	34	15



c_{ij}	y_1	y_2	y_3	y_4	y_5	y_6
x_1	12	10	31	27	10	30
x_2	22	7	25	15	11	14
x_3	19	7	19	10	15	15
x_4	10	6	21	19	14	24
x_5	15	23	14	24	31	34
x_6	35	26	16	9	34	15

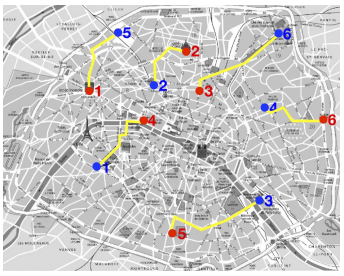
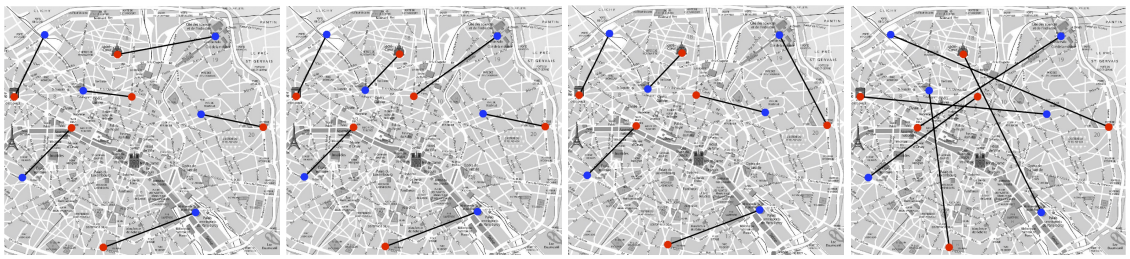


Figure 1: Cost matrix and associated connections. Left: a row of the cost matrix. Right: a particular example of permutation.



Cost=64 Cost=65 Cost=66 Cost=152

Figure 2: Examples of permutations with different costs.

such a choice of connections. Figure 1 illustrates in the center and on the right the example

$$\sigma(1) = 5,\ \sigma(2) = 2,\ \sigma(3) = 6,\ \sigma(4) = 1,\ \sigma(5) = 3,\ \sigma(6) = 4. \tag{1}$$

The mass conservation constraint means that σ is a bijection of the set $\{1, \dots, 6\}$ within itself. We also say that σ is a permutation.

The transport cost associated with such a bijection is the sum of the costs $C_{i,\sigma(i)}$ selected by the permutation σ, that is to say

$$\text{Cost}(\sigma) \stackrel{\text{def.}}{=} C_{1,\sigma(1)} + C_{2,\sigma(2)} + C_{3,\sigma(3)} + C_{4,\sigma(4)} + C_{5,\sigma(5)} + C_{6,\sigma(6)}. \tag{2}$$

For example, for the bijection (1) shown in Figure 1, we obtain as cost

$$C_{1,5} + C_{2,2} + C_{3,6} + C_{4,1} + C_{5,3} + C_{6,4} = 10 + 7 + 15 + 10 + 14 + 9 = 65.$$

Monge's problem is to look for the permutation σ which has the minimum cost, that is to solve the optimization problem

$$\min_{\sigma \in \Sigma_6} \text{Cost}(\sigma), \tag{3}$$

where we noted Σ_6 the set of permutations of the set $\{1, \dots, 6\}$.

Figure 2 shows that the permutation (1) is not the best: there exists for example another permutation which has a cost of 64. But is this the best? It turns out that it is the case, since we can indeed test on a computer all the permutations of $\{1, \dots, 6\}$ and calculate their cost. How many permutations are there in total? To make this count, we see that there are six possible assignment choices from 1 to $\sigma(1) \in \{1, \dots, 6\}$, then five possible choices to assign 2 to $\sigma(2) \in$

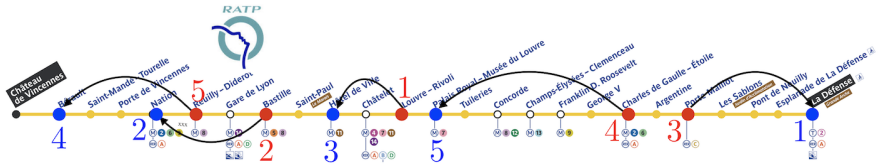


Figure 3: Optimal transport in 1D along a metro line. The optimal bijection is $\sigma : (1, 2, 3, 4, 5) \mapsto (3, 2, 1, 5, 4)$.

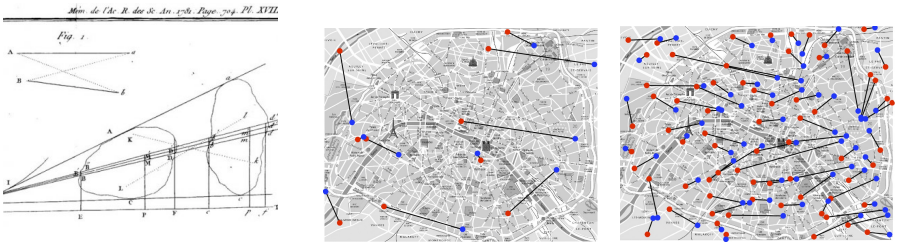


Figure 4: Left: excerpt from Monge's article [6]. Right: the optimal transport in 2D for a Euclidean cost.

$\{1, \ldots, 6\} - \{\sigma(1)\}$, and so on. The total number of possibilities is thus $6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$ that we note 6!, "factorial 6". If we consider a number n of bakeries, then the number of permutations to test to find the best is $n! = n \times (n-1) \times \ldots \times 2 \times 1$. This number grows extremely fast with n, for example $70! \approx 1.198 \times 10^{100}$, to be compared with the 10^{11} neurons in the brain and the 10^{79} atoms in the universe. This exhaustive search strategy is only possible for very small values of n.

1.2 In 1D and 2D

Section 2 explains how mathematical advances have made it possible to develop efficient techniques for calculating an optimal transport σ even for large values of n. But it took almost 200 years to get there. In some simple cases, however, the optimal transport can be calculated in a simple way. The most basic case is when the points to be matched are along a 1D axis, for example if cafés and bakeries are located along a subway line. It is also necessary that the cost $C_{i,j}$ be the distance along this axis (eg the metro travel time between the stations). In this case, simply rank the indices i and j in ascending order (thus from left to right along the subway line) and match the first index i to the first index j together, then the second index, etc. This process is illustrated in Figure 3. The calculation time required to calculate the optimal transport by subway is therefore the time required to classify the indices. The simplest algorithm for ranking is the one usually used to sort a set of n cards: it is the insertion sort, which iteratively inserts each card in its place relative to the cards already classified. It performs $n(n-1)/2$ comparisons. For $n = 70$, this requires only 21415 operations, which makes the method usable, unlike the exhaustive search of all $n!$ Permutations. There are even faster algorithms (eg merge sort), which perform on the order of $n \log(n)$ operations, and hence for $n = 70$, such methods require less than 1000 operations.

Unfortunately, it is no longer possible to use this sorting technique in more general cases. For points in dimension 2, if we take as cost $C_{i,\sigma(i)}$ the Euclidean distance (the flight distance) between the points, then Gaspard Monge showed in his paper original (see Figure 4, left) that optimal transport can not contain crossover. For example, as shown in Figure 4 (on the right), if we trace

all the segments between the points $i \mapsto j = \sigma(i)$ that the we connect by the bijection defined by an optimal σ, these never cross each other. This geometrical observation is however not sufficient to compute an optimal transport in 2D: there are indeed many permutations σ such that the associated segments do not intersect. It will be necessary to analyze more finely the structure of the optimal permutations in order to be able to calculate them in an efficient way. We will now see how Leonid Kantorovich has reformulated the problem of Monge in order to achieve this.

2 Optimal Transport of Kantorovich

Leonid Kantorovitch is a Soviet mathematician and economist who revolutionized the theory of optimal transport during the 1940s. His research stemmed from practical considerations that occupied him before and after the Second World War. He played an important role in ensuring an optimal distribution of resources, especially during the Leningrad siege. At the same time, he has been involved in the development of modern optimization, which has had a huge impact in a large number of applied fields. He obtained the Nobel Prize in Economics in 1975, because the first applications (but certainly not the only!) of his theory have been in this field.

2.1 Kantorovich's Problem

Kantorovich's central idea is to modify Monge's problem by replacing the set of permutations with a larger but simpler set. First we notice that we can represent a permutation $\sigma \in \Sigma_n$ using a permutation matrix P which is a binary matrix (filled with 0 and 1) of size $n \times n$ such that $P_{i,j} = 0$ unless $j = \sigma(i)$ in which case $P_{i,\sigma(i)} = 1$. For example, for $n = 3$ points, the permutations $(1,2,3) \mapsto (1,2,3)$ (the identity), $(1,2,3) \mapsto (3,2,1)$ and $(1,2,3) \mapsto (2,1,3)$ are represented by size 3×3 matrices

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

In the following, $\mathcal{P}_n$ is the set of $n!$ Permutation matrices of size $n \times n$.

Since the matrix is binary, with only n non-zero elements equal to 1, we can replace the sum of n terms that appears in $\mathrm{Cost}(\sigma)$ defined in (2) by a sum over the set of $n \times n$ indices (i,j), that is, if P is the permutation matrix associated with σ, we have

$$\mathrm{Cost}(\sigma) = \sum_{i=1}^{n} \sum_{j=1}^{n} P_{i,j} C_{i,j}.$$

We can thus replace the problem of Monge (3) by the equivalent problem

$$\min_{P \in \mathcal{P}_n} \sum_{i=1}^{n} \sum_{j=1}^{n} P_{i,j} C_{i,j}. \tag{4}$$

Kantorovich's genius has been to remark that we can replace the discrete set $\mathcal{P}_n$ (that is to say composed of a finite, but very large, set of $n!$ Matrices) by a set which is "continuous" (so in particular infinite) but which is simpler. Note that the permutation matrices of $\mathcal{P}_n$ are exactly the matrices that have one and only one along each row and column. This can also be expressed as the

fact that a permutation matrix is a binary matrix whose sum of each row and of each column is 1, that is to say

$$\mathcal{P}_n = \left\{ P \in \{0,1\}^{n\times n} \;;\; \forall\, i, \sum_j P_{i,j} = 1, \forall\, j, \sum_i P_{i,j} = 1 \right\}.$$

What makes this set very complicated is the binary constraint, that is, these matrices are constrained to be in $\{0,1\}^{n\times n}$. Kantorovitch then proposes to "relax" this constraint by simply assuming that the entries of P are between 0 and 1. This defines a larger set, the set of bistochastic matrices

$$\mathcal{B}_n \overset{\text{def.}}{=} \left\{ P \in [0,1]^{n\times n} \;;\; \forall\, i, \sum_j P_{i,j} = 1, \forall\, j, \sum_i P_{i,j} = 1 \right\}. \tag{5}$$

The Kantorovitch problem is obtained by performing this replacement in (4), in order to solve

$$\min_{P\in\mathcal{B}_n} \sum_{i=1}^n \sum_{j=1}^n P_{i,j} C_{i,j}. \tag{6}$$

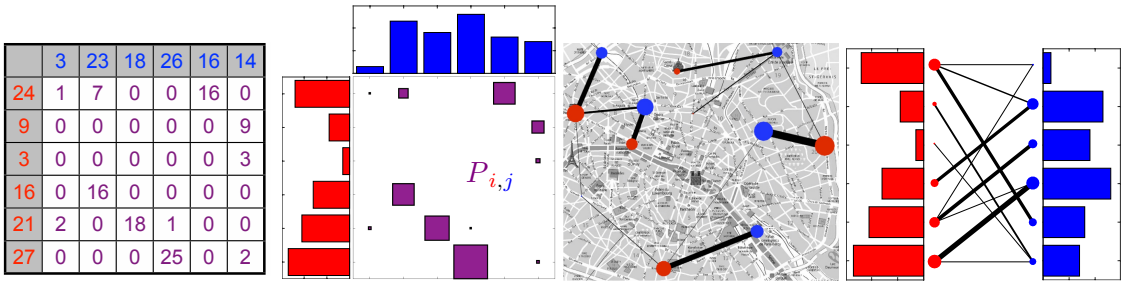
The huge advantage of the Kantorovich (6) problem over that of Monge (4) is that the set of bistochasic matrices is convex, ie if we consider two bistochasic matrices $P, Q \in \mathcal{B}_n$, so their mean $\frac{P+Q}{2} \in \mathcal{B}_n$ is still bistochasic. This is not true for permutation matrices, since the average of two binary matrices (P, Q) is not binary (except of course if $P = Q$). This convexity is the key to the development of efficient algorithms. This new formulation has indeed benefited from a second revolution initiated by George Dantzig [4], which, at the same time, proposed the algorithm of the simplex. This one allows to solve efficiently a certain class of convex optimization problems: linear programming problems, of which (6) is a particular case. In the case of the Kantorovitch problem, there is indeed a simplex algorithm that has a complexity of the order of n^3 operations, which allows calculations to be made for large n, of the order several thousands.

2.2 Monge–Kantorovich Equivalence

The set of bistochastic matrices is larger than the set of permutations matrices, $\mathcal{P}_n \subset \mathcal{B}_n$, so that we have the inequality

$$\min_{P\in\mathcal{B}_n} \sum_{i=1}^n \sum_{j=1}^n P_{i,j} C_{i,j} \leqslant \min_{P\in\mathcal{P}_n} \sum_{i=1}^n \sum_{j=1}^n P_{i,j} C_{i,j} \tag{7}$$

between the problems of Kantorovich and Monge. But a fundamental theorem due to George Birkhoff and John von Neumann [2, 10] ensures that in fact there is equality between the values of these two minimizations. Indeed, this theorem shows that there is always a solution matrix of the Kantorovitch problem which is a matrix of permutation, so that it is also a solution to Monge's problem. Beware however, in general there is no uniqueness of the solutions of these problems: there may exist a bistochastic matrix solution of the Kantorovich problem which is not a permutation. The combination of two spectacular advances, due to Kantorovich and Dantzig, made optimal transport applicable to large scale problems, since the simplex algorithm can be used to solve these problems in practice.



(a) matrix (b) histograms (c) segments (d) bipartite graph

Figure 5: Different ways of representing a coupling matrix $P \in \mathcal{B}(a,b)$: (a) a table of numbers whose rows and columns have prescribed sums; (b) a two-dimensional histogram whose square size is propositional to $P_{i,j}$; (c) a set of segments whose width is proportional is $P_{i,j}$. (d) a bipartite graph, that is to say with two sets of vertices such that the edges are only between these two sets.

2.3 The Weighted Case

In addition to its practical interest, Kantorovich's formulation has also allowed generalizing Monge's initial problem, by giving the right framework to formalize it and study it mathematically. Indeed, Monge's problem is quite limited. What happens for example if there is not the same number n café and m bakeries? The initial problem (3) has no solution, because you can not put in bijection two sets of different sizes. The right concept is not the number of bakeries and cafés, but rather the $(a_1,\dots,a_n)$ production distributions (associated with bakeries) and the $(b_1,\dots,b_m)$ of café consumption. For example, if the first bakery produces 45 croissants a day, we will take $a_1 = 45$, and $b_3 = 34$ means that the 3^{rd} café consumes 34 croissants a day. In the case initially considered, where $n = m$, all the quantities a_i and b_j are equal to 1. But in many concrete cases, these quantities are arbitrary. These quantities must be positive, and satisfy

$$a_1 + \cdots + a_n = b_1 + \cdots + b_m,$$

so that there is as much production as consumption. Kantorovich's construction naturally adapts to this case of general distributions, replacing the bistochasic matrices (5) by matrices of "coupling" which satisfy the mass conservation constraint.

$$\mathcal{B}(a,b) \overset{\text{def.}}{=} \left\{ P \in [0,1]^{n\times m} \;;\; \forall\, i, \sum_j P_{i,j} = a_i, \forall\, j, \sum_i P_{i,j} = b_j \right\}.$$

In the original case $n = m$ and $a_i = b_j = 1$, then $\mathcal{B}(a,b) = \mathcal{B}_n$ which corresponds to doubly stochastic matrices. In the general case, whenever an entry $P_{i,j}$ is non-zero, this means that there is some transfer of "mass" (here a certain amount of croissants) between i and j. As shown in Figure 5, we can visualize in different ways such a matrix P coupling two distributions (a,b). Unlike the case of doubly stochastic matrix, for which there is always a solution that is a permutation, here optimum coupling $\mathcal{B}(a,b)$ can have more than one non-zero entry $P_{i,j}$ along a line indexed by i (see Figure 5). This means that this bakery i is connected to several cafés, so that its production is then separated into several batches distributed while meeting conservation constraint of the mass $\sum_j P_{ij} = a_i$.

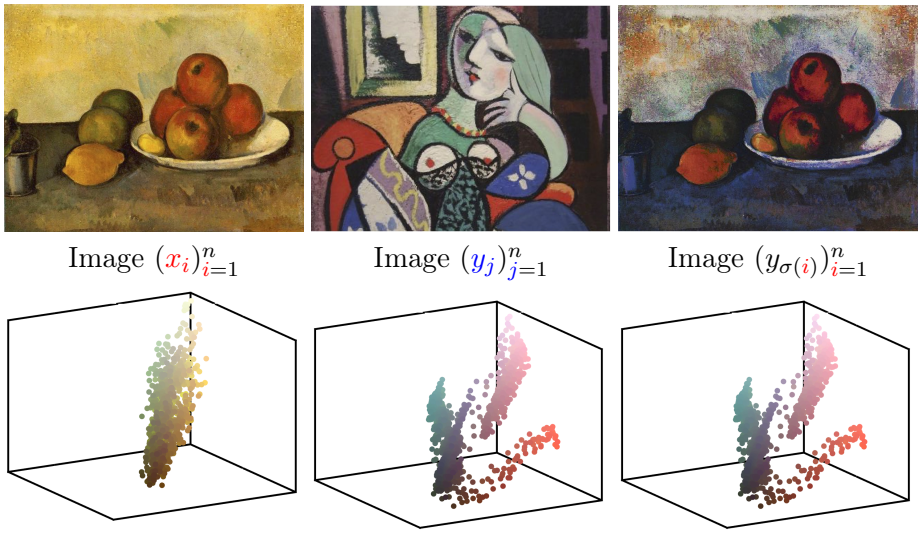


Figure 6: Example of transferring color palettes using optimal transport. Top: The pixels are on the display grid to form a color image. Bottom: The pixels are placed at their positions in $\mathbb{R}^3$ to form a scatter plot.

Kantorovich's problem which generalizes (6) is then written

$$\min_{P \in \mathcal{B}(a,b)} \sum_{i=1}^{n} \sum_{j=1}^{m} P_{i,j} C_{i,j} \tag{8}$$

which means that you have to pay $C_{i,j}$ each time you transfer a unit of mass between i and j. Just like the original problem (6), we can solve it effectively with the simplex algorithm. Figure 5 shows an example of optimal coupling.

3 Applications

Although the initial motivations of Monge and Kantorovitch were respectively military and economic, the optimal transport finds countless applications, both theoretical but also more concrete. Mathematically, one can consider "continuous" distributions of masses, somehow the limit when the number of point n tends to infinity. This makes it possible to define the transport problem between any probability measurements. This theoretical point of view is extremely fruitful, and it was the French mathematician Yann Brenier who first showed equivalence in the continuous framework of the formulations of Monge and Kantorotich [3]. These pioneering works showed the connection between the transport problem and the partial differential equations, and led, among other things, to the Fields medals of Cédric Villani (2010) and Alessio Figalli (2018).

Optimal transport has recently become the focus of more applied problems in data sciences, especially to solve problems in image processing and machine learning. The first idea, the most immediate, is to use the bijection σ to transform data, for example images. In this case, the pixels $(x_i)_{i=1}^n$ and $(y_j)_{j=1}^n$of two color images are considered. Each pixel $x_i, y_j \in \mathbb{R}^3$ is a vector of dimension 3, which represents the intensities of each of the three elementary colors, red, green and blue. In order to change the colors of the first image, and impose the palette of the second image, we calculate

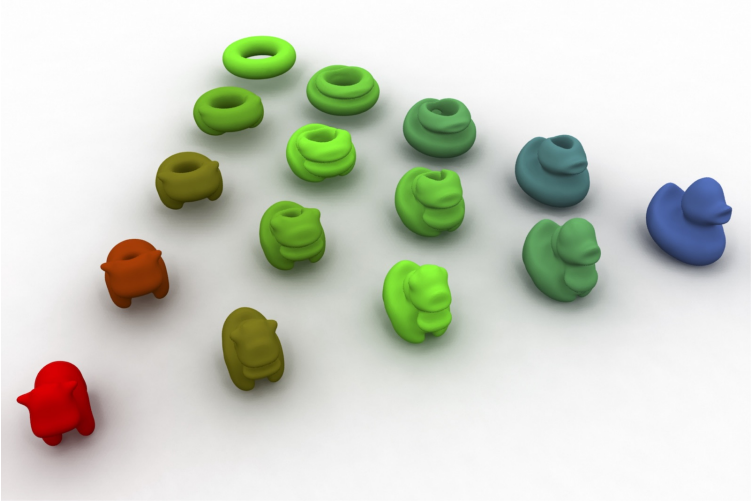


Figure 7: Example of barycentric interpolation between 3D forms, obtained by minimizing (9).

the transport σ for the cost matrix $C_i = \|x_i - y_j\|^2$ (that is, the square of the Euclidean norm in $\mathbb{R}^3$), which is the square of the Euclidean distance between the pixels. The image with the modified colors is $(y_{\sigma(i)})_{i=1}^n$, ie we replace in the first image the pixel x_i by the pixel $y_{\sigma(i)}$. This image looks like the first, but has the color palette of the second image. Figure 6 illustrates this process to impose the color palette of Picasso to a painting by Cézanne.

Optimal transport can also be used for more difficult problems, by only indirectly using the bijection σ or the optimal coupling matrix $P \in \mathcal{B}(a, b)$. The central idea is that the quantity associated with an optimal coupling P solution of (8)

$$W(a, b) \stackrel{\text{def.}}{=} \sum_{i,j} P_{i,j} C_{i,j}$$

somehow defines the effort required to move the mass of the a distribution to the b distribution. It allows to quantify how much these two distributions are "close". For example, if $C_{i,j} = \|x_i - y_j\|^2$ is the square of the Euclidean distance between points, then the quantity $W(a, b)^{1/2}$ is a distance between the distributions, in particular it satisfies $W(a, b) = 0$ if and only if $a = b$, and it satisfies the inequality triangular. These properties are very important for applying transport to practical problems.

A typical example of the application of this W quantity is to compute centroids between [1] distributions. Figure 7 shows an example where we consider three distributions a, b, c (shown at the three vertices of the triangles) which are uniform mass distributions inside 3D shapes (that is, the mass a_i associated with the i^{th} point is 0 outside the first form and takes a constant value inside). A weighted barycenter of these three distributions is calculated by mimicking the fact that in a Euclidean space, the weighted centroid r of three points x, y, z minimizes the sum of distances squared.

$$\min_r \alpha\|x - r\|^2 + \beta\|y - r\|^2 + \gamma\|z - r\|^2,$$

where the weights (α, β, γ) are the weightings of the centroid, which are positive reals and such that $\alpha + \beta + \gamma = 1$. The weighted barycenter d of (a, b, c) thus minimizes the weighted sum of optimal

Figure 8: Examples of histograms of word distributions in two different texts (only the most frequent words are shown).

transport distances

$$\min_d \alpha W(a,d) + \beta W(b,d) + \gamma W(c,d). \tag{9}$$

By modifying the weights (α, β, γ), we modify the shape obtained by moving inside an optimal transport triangle. This W distance can be used for many other applications where probability distributions must be compared. This is the case in machine learning, for example to compare texts using the distributions of words that compose them. Figure 8 illustrates the histograms of appearance of words for two texts, where the size of the letters of the word i is proportional to the mass a_i. A difficult question in this case is which cost matrix $C_{i,j}$ to use between two words (i, j). It is a linguistics problem (to characterize the semantic proximity between words of a language), which one can seek to solve at the same time as the optimal transport [5].

Conclusion

Optimal transport has seen many revolutions. Led by mathematicians such as Monge, Kantorovich, Danzig and Brenier, it has gradually become a fundamental theoretical and numerical tool. It is now at the heart of important questions in data science to model, numerically solve and theoretically analyze problems in machine learning. The opportunities to develop new theories and powerful algorithms are immense. For more information on the theoretical aspects of optimal transport, one can consult the books [9, 8]. Numerical and applicative aspects are covered in the book [7].

Acknowledgments

I would like to thank Vincent Beck, Gwenn Guichaoua and Marie-Noëlle Peyré for their careful proofreadings.

References

[1] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] Garrett Birkhoff. Tres observaciones sobre el algebra lineal. *Universidad Nacional de Tucumán Revista Series A*, 5:147–151, 1946.

[3] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

[4] George B Dantzig. Application of the simplex method to a transportation problem. *Activity Analysis of Production and Allocation*, 13:359–373, 1951.

[5] Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. Supervised word mover's distance. In *Advances in Neural Information Processing Systems*, pages 4862–4870, 2016.

[6] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pages 666–704, 1781.

[7] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *to appear in Fundation and Trends in Machine Learning*, 2018.

[8] Filippo Santambrogio. *Optimal transport for applied mathematicians*. Birkhauser, 2015.

[9] Cedric Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics Series. American Mathematical Society, 2003.

[10] John Von Neumann. A certain zero-sum two-person game equivalent to the optimal assignment problem. *Contributions to the Theory of Games*, 2:5–12, 1953.