An Introduction to Data Sciences

Gabriel Peyré
CNRS & DMA, École Normale Supérieure, PSL

CHAPTER 4

Optimal transport

September 9, 2026

Comparing two distributions requires a notion of the cost of moving mass between locations. Optimal transport turns that idea into an optimization problem. Starting with a small assignment problem, we introduce transport plans, explain when they coincide with maps, and use transport distances to compare colors, shapes, and texts.

4.1 Monge's problem

In 1781, Gaspard Monge formulated the problem of moving earth from one place to another at minimum cost [7]. His cost was the distance traveled by each unit of mass. Other applications suggest other costs, such as travel time or the squared Euclidean distance.

4.1.1 An assignment problem

Consider six bakeries and six cafés in Paris. Each bakery supplies one batch of croissants, and each café requires one batch. In the assignment model, each batch goes to exactly one café, and each café receives exactly one batch. Both requirements matter: merely giving every bakery a destination would allow several bakeries to choose the same café.

Let C_{ij} be the travel time from bakery i to café j. The example in Figure 4.1 has $C_{34} = 10$. An admissible assignment is a permutation σ of $\{1, \dots, 6\}$. For instance,

$$(\sigma(1), \dots, \sigma(6)) = (5, 2, 6, 1, 3, 4). \tag{4.1}$$

Its total cost is

$$\mathrm{Cost}(\sigma) = \sum_{i=1}^{6} C_{i,\sigma(i)} = 10 + 7 + 15 + 10 + 14 + 9 = 65. \tag{4.2}$$

This adds the six journey times. It does not model a single vehicle making a tour, nor the completion time of parallel deliveries.

For n equal batches on each side, the discrete Monge problem is

$$\min_{\sigma \in \Sigma_n} \sum_{i=1}^{n} C_{i,\sigma(i)}, \tag{4.3}$$

where Σ_n is the set of permutations. In our example, the optimal value is 64, achieved by $(5, 6, 2, 1, 3, 4)$. Exhaustively checking the $6! = 720$ possibilities verifies this claim. There are $n!$ assignments in general, and $70! \approx 1.198 \times 10^{100}$, so enumeration quickly becomes impractical.

4.1.2 Sorting in one dimension

Suppose the source locations $x_1 \leq \cdots \leq x_n$ and target locations $y_1 \leq \cdots \leq y_n$ lie on a line, with cost $C_{ij} = |x_i - y_j|$. Matching the points in sorted order is optimal. The key inequality is

$$|x - y| + |x' - y'| \leq |x - y'| + |x' - y| \quad \text{when } x \leq x' \text{ and } y \leq y'.$$

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

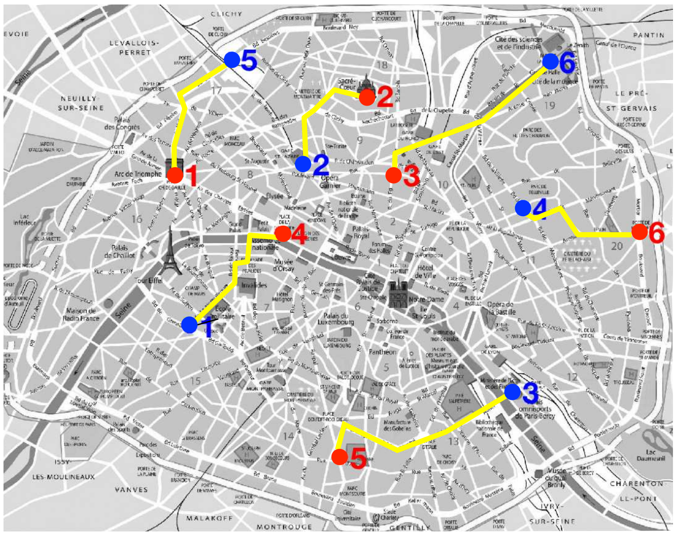

Figure 4.1. Cost matrix and an assignment of the six bakeries (red) to the six cafés (blue). The highlighted entries give a total journey time of 65 minutes.

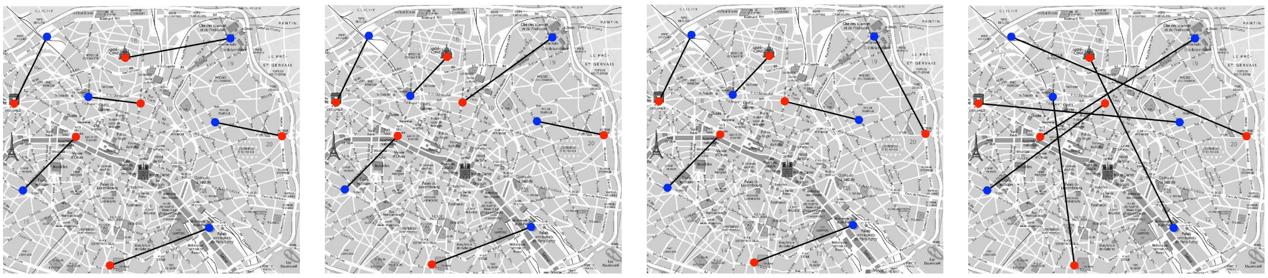

Figure 4.2. Four assignments with costs 64, 65, 66, and 152, from left to right.

One can verify it by ordering the four points on the line. Whenever a permutation contains an inversion, swapping the two assigned targets does not increase the cost. Removing inversions successively produces the sorted matching. Equality can occur, so this optimal assignment need not be unique.

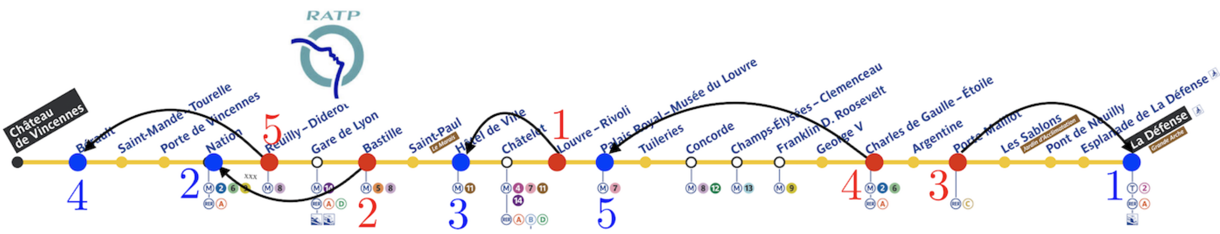

Figure 4.3. Sorted matching along a metro line, considered as a one-dimensional axis. For the indices shown, $\sigma : (1,2,3,4,5) \mapsto (3,2,1,5,4)$.

The computational cost is therefore that of sorting the two lists. Insertion sort uses at most $n(n-1)/2$ comparisons per list, which is 2,415 for $n = 70$. Merge sort reduces this to order $n \log n$. These are comparison counts for sorting, rather than estimates of every operation needed by a transport application.

In two dimensions there is no single ordering that solves the general problem. For the Euclidean distance cost, however, two segments in an optimal matching cannot cross transversely in their interiors. If segments $[x, y]$ and $[x', y']$ meet at z, the triangle inequality gives

$$|x - y'| + |x' - y| \le |x - z| + |z - y'| + |x' - z| + |z - y| = |x - y| + |x' - y'|.$$

For a transverse crossing the inequality is strict, so exchanging destinations decreases the cost. Collinear overlaps or shared endpoints are degenerate cases and need not give a strict improvement. This observation does not suffice to compute an optimum: many noncrossing assignments can have different costs.

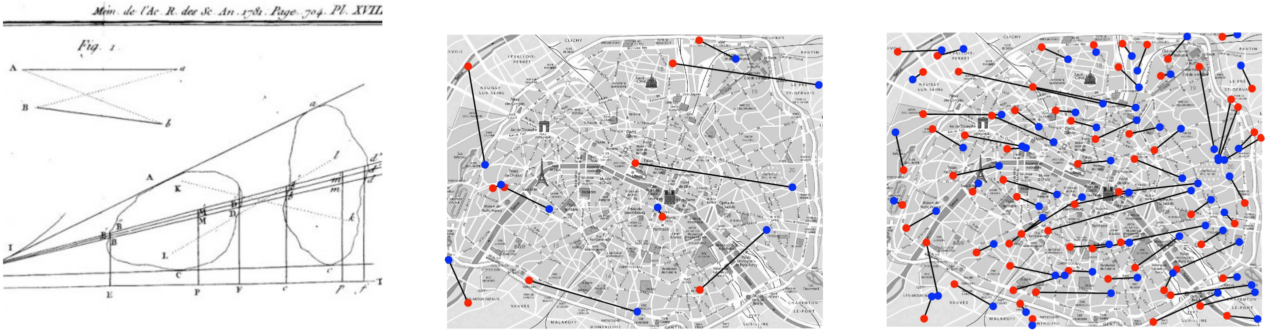

Figure 4.4. An illustration from Monge's original article [7] and two optimal Euclidean matchings in the plane.

4.2 Kantorovich's relaxation

Kantorovich's formulation allows mass to split between destinations [6]. It replaces discrete choices by nonnegative quantities subject to linear conservation constraints, making optimal transport a linear programming problem.

4.2.1 From permutations to matrices

Represent a permutation by $P_{ij} = 1$ when $j = \sigma(i)$ and $P_{ij} = 0$ otherwise. For example, the permutations $(1, 2, 3)$, $(3, 2, 1)$, and $(2, 1, 3)$ correspond to

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Each row and each column sums to one. If $\mathcal{P}_n$ denotes this set of permutation matrices, then (4.3) is equivalent to

$$\min_{P \in \mathcal{P}_n} \langle C, P \rangle, \qquad \langle C, P \rangle = \sum_{i,j} C_{ij} P_{ij}. \tag{4.4}$$

Relaxing the binary constraint gives the doubly stochastic matrices

$$\mathcal{B}_n = \left\{ P \in \mathbb{R}_+^{n \times n} : \sum_j P_{ij} = 1, \ \sum_i P_{ij} = 1 \right\}. \tag{4.5}$$

Nonnegativity and the row sums already imply $P_{ij} \le 1$. The relaxed problem is

$$\min_{P \in \mathcal{B}_n} \langle C, P \rangle. \tag{4.6}$$

The set $\mathcal{B}_n$ is compact and convex: for $P, Q \in \mathcal{B}_n$ and $0 \le t \le 1$, the matrix $(1-t)P + tQ$ is still feasible. A minimizer exists because the objective is continuous.

The development of linear programming, including Dantzig's simplex method [4], made these problems computationally accessible. Specialized assignment algorithms, such as implementations of the Hungarian method, solve the dense $n \times n$ assignment problem in $O(n^3)$ operations. This bound should not be attributed to the general simplex method, whose worst-case running time is not polynomial.

4.2.2 Why the relaxation is exact for equal masses

The Birkhoff–von Neumann theorem states that every doubly stochastic matrix is a convex combination of permutation matrices [2, 11]:

$$P = \sum_{k=1}^{K} \lambda_k P^{(k)}, \qquad \lambda_k \ge 0, \quad \sum_k \lambda_k = 1.$$

Consequently, $\langle C, P \rangle = \sum_k \lambda_k \langle C, P^{(k)} \rangle$ cannot be smaller than the least permutation cost. Since every permutation matrix is feasible for the relaxation, we obtain

$$\min_{P \in \mathcal{B}_n} \langle C, P \rangle = \min_{P \in \mathcal{P}_n} \langle C, P \rangle. \tag{4.7}$$

Thus at least one optimal plan is a permutation. It does not follow that every optimal plan is a permutation: the average of two distinct optimal permutations is also optimal. If all costs vanish, every doubly stochastic matrix is optimal.

A useful way to certify an optimal value is to find numbers u_i, v_j satisfying $u_i + v_j \le C_{ij}$. Every $P \in \mathcal{B}_n$ then satisfies

$$\langle C, P\rangle \ge \sum_{i,j}(u_i + v_j)P_{ij} = \sum_i u_i + \sum_j v_j.$$

For the Paris example, $u = (2, 3, 3, 0, 5, 2)$ and $v = (10, 4, 9, 7, 8, 11)$ satisfy all 36 inequalities and their entries sum to 64. Together with an assignment of cost 64, they prove optimality without listing all permutations.

4.2.3 Different numbers of points and unequal masses

Now let n bakeries supply amounts $a_i \ge 0$, and let m cafés require amounts $b_j \ge 0$, with common total mass

$$\sum_{i=1}^{n} a_i = \sum_{j=1}^{m} b_j = s.$$

A transport plan belongs to

$$\mathcal{B}(a, b) = \left\{ P \in \mathbb{R}_+^{n\times m} : \sum_j P_{ij} = a_i,\ \sum_i P_{ij} = b_j \right\}.$$

When $s > 0$, the plan $P_{ij} = a_i b_j / s$ proves feasibility. When $s = 0$, the only plan is zero. The transport problem is

$$\min_{P\in\mathcal{B}(a,b)} \sum_{i=1}^{n}\sum_{j=1}^{m} C_{ij}P_{ij}. \tag{4.8}$$

For finite costs, compactness again ensures the existence of a minimizer. An entry P_{ij} is the amount shipped from i to j, and C_{ij} is its cost per unit of mass.

	3	23	18	26	16	14
24	1	7	0	0	16	0
9	0	0	0	0	0	9
3	0	0	0	0	0	3
16	0	16	0	0	0	0
21	2	0	18	1	0	0
27	0	0	0	25	0	2

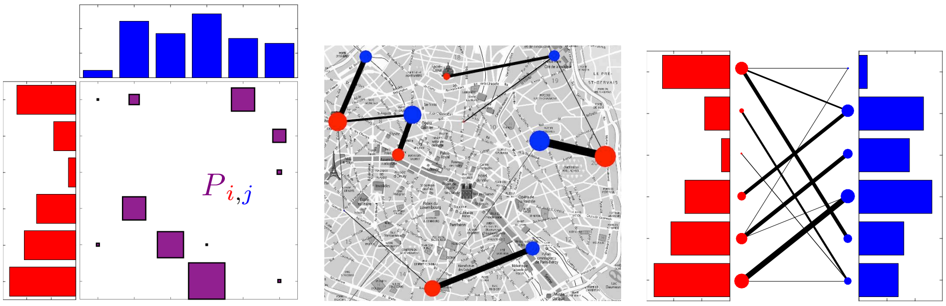

Figure 4.5. Four views of the same coupling: its matrix entries, a display of their magnitudes by squares, weighted segments in space, and a bipartite graph. Row and column sums give the prescribed masses a and b.

Mass splitting can now be necessary. One bakery with mass 2 and two cafés requiring mass 1 each have the unique feasible plan $P = (1\ 1)$. A map that sends the entire source mass to a single target cannot realize these marginals. This explains why a transport plan is more general than a transport map. In the equal-mass case $a_i = b_j = 1$, we recover $\mathcal{B}_n$; for probability masses $a_i = b_j = 1/n$, the corresponding plans are P/n.

4.3 Measures and transport distances

The same ideas apply to probability measures, including discrete distributions $\mu = \sum_i a_i \delta_{x_i}$ and $\nu = \sum_j b_j \delta_{y_j}$. A coupling $\pi \in \Pi(\mu, \nu)$ is a probability measure on pairs (x, y) with marginals μ and ν. A measurable map T induces a coupling $(\mathrm{Id}, T)_{\#}\mu$ when $T_{\#}\mu = \nu$; the notation means that applying T to a random variable of law μ gives one of law ν.

For probability measures on $\mathbb{R}^d$ with finite second moments, the quadratic transport distance is

$$W_2(\mu, \nu)^2 = \min_{\pi\in\Pi(\mu,\nu)} \int |x - y|^2 \,\mathrm{d}\pi(x, y). \tag{4.9}$$

In the discrete case, this is (4.8) with $C_{ij} = |x_i - y_j|^2$. The distance depends on both masses and locations: equal weight vectors at different locations need not describe the same measure. The square root W_2 is a metric, and $W_2(\mu, \nu) = 0$ exactly when $\mu = \nu$. Its square generally fails the triangle inequality; for point masses at 0, 1, and 2, the relevant squared distances are 1, 1, and 4.

Brenier's theorem gives a precise condition under which the optimal coupling is a map [3]. If the source μ is absolutely continuous with respect to Lebesgue measure and both measures have finite second moments, the quadratic problem has a unique optimal coupling. It is induced by a map $T = \nabla\varphi$, where φ is convex, and T is unique μ-almost everywhere. Absolute continuity and the choice of cost matter; one cannot assert this conclusion for arbitrary measures and arbitrary costs.

4.4 Applications

4.4.1 Transferring a color palette

For two images with the same number n of pixels, let $x_i, y_j \in \mathbb{R}^3$ be their RGB colors. An assignment minimizing $\sum_i |x_i - y_{\sigma(i)}|^2$ gives an output whose color at source position i is $y_{\sigma(i)}$. Every target color is used once, so the output has exactly the target's empirical color distribution. The optimization limits changes in color, but contains no spatial regularity term; preservation of edges and textures is therefore not guaranteed.

Figure 4.6. Color-palette transfer. Columns: source, target, and recolored source. Top: colors on the image grid; bottom: the corresponding point clouds in RGB space. A permutation of the target colors gives the same color cloud in the last two columns.

This distinction separates color transport from moving pixels in the image plane. Here the locations being transported are points in color space; the display grid stays fixed. With different image sizes or weighted color clusters, one uses a coupling and a specified rule to convert that coupling into output colors. A barycentric average of coupled colors need not preserve the target histogram exactly.

4.4.2 Barycenters of shapes

The Euclidean weighted mean of three points minimizes $\alpha|x - r|^2 + \beta|y - r|^2 + \gamma|z - r|^2$, where $\alpha, \beta, \gamma \geq 0$ and $\alpha + \beta + \gamma = 1$. Its value is $r = \alpha x + \beta y + \gamma z$. Replacing squared Euclidean distances by squared transport distances defines a Wasserstein barycenter [1]:

$$\min_{\rho \in \mathcal{P}_2(\mathbb{R}^d)} \alpha W_2(\mu, \rho)^2 + \beta W_2(\nu, \rho)^2 + \gamma W_2(\eta, \rho)^2. \tag{4.10}$$

Here $\mathcal{P}_2$ is the set of probability measures with finite second moment. In a grid-based computation, one instead restricts ρ to probability weights on that grid. A shape can be represented by a uniform density inside it, normalized to total mass one. Changing the three weights produces a family of interpolated distributions (Figure 4.7). These distributions need not themselves be uniform inside a sharp shape, and uniqueness requires additional assumptions.

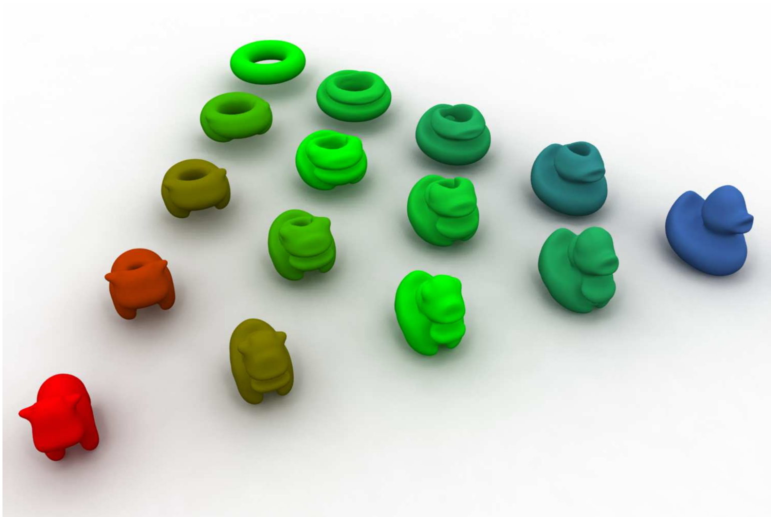

Figure 4.7. Numerical barycentric interpolation of three-dimensional shapes represented by probability distributions. The three reference shapes occupy the vertices of the triangle; interior points correspond to different weights in (4.10).

4.4.3 Comparing texts

Normalize word counts in a text to form a probability distribution. Transport can compare two such distributions using a cost that describes the dissimilarity between words, for example distances between word embeddings. The choice of this ground cost determines which substitutions are considered inexpensive; it may also be learned from a task [5]. Word histograms discard word order, so they cannot by themselves capture all grammatical or semantic distinctions.

Figure 4.8. Word clouds showing the most frequent words in two texts. Their sizes indicate relative frequencies; these are illustrations rather than a complete numerical histogram.

Further reading. For the theory, see [9, 10]. Computational methods and applications are developed in [8]. The companion course Optimal Transport for Machine Learning (OT4ML) offers another route into these ideas.

Acknowledgements. I thank Vincent Beck, Gwenn Guichaoua, and Marie-Noëlle Peyré for their careful reading.

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 álgebra 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. In Tjalling C. Koopmans, editor, *Activity Analysis of Production and Allocation*, pages 359–373. Wiley, 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] Leonid Kantorovich. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 37(2):227–229, 1942.

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

[8] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. URL https://arxiv.org/abs/1803.00567.

[9] Filippo Santambrogio. *Optimal transport for applied mathematicians*. Birkhäuser, 2015.

[10] Cédric Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics Series. American Mathematical Society, 2003. ISBN 9780821833124.

[11] 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.