

# Le transport optimal numérique et ses applications

Gabriel Peyré  
CNRS & DMA  
École Normale Supérieure  
[gabriel.peyre@ens.fr](mailto:gabriel.peyre@ens.fr)

## Résumé

Le transport optimal est un problème très ancien, formulé par Monge au 18<sup>e</sup> siècle. Il a cependant fallu plusieurs révolutions mathématiques pour qu'il devienne un outil incontournable à la fois en théorie et en pratique. Cet article retrace ces révolutions, initiées par Leonid Kantorovitch pendant la seconde guerre mondiale. La formulation qu'il a proposée se prête à une analyse mathématique poussée ainsi qu'à son application à de nombreux problèmes. Elle fait du transport optimal un outil de choix pour aborder l'explosion récente de la science des données.

## 1 Le Transport Optimal de Monge

Gaspard Monge, en plus d'être un grand mathématicien, a participé activement à la révolution Française, et a créé l'École Polytechnique ainsi que l'École Normale Supérieure. Motivé par des applications militaires, il a formulé en 1781 le problème du transport optimal [7] : il s'est posé la question du calcul de la façon la plus économique de transporter de la terre entre deux endroits pour faire des remblais. Dans son texte original, il a fait l'hypothèse que le coût du déplacement d'une unité de masse est égal à la distance parcourue, mais on peut utiliser n'importe quel coût adapté au problème à résoudre.

### 1.1 Le problème de Monge

Pour illustrer le problème et sa formulation mathématique, intéressons-nous à la façon optimale de distribuer les croissants depuis les boulangeries vers les cafés, le matin dans Paris. Pour simplifier, nous allons supposer qu'il y a uniquement six boulangeries et cafés, que l'on peut voir à la figure 1 (les boulangeries sont en rouge et les cafés en bleu). On suppose que toutes les boulangeries produisent le même nombre de croissants et que tous les cafés demandent également ce même nombre de croissant. Le coût à minimiser est le temps total des trajets, et l'on note  $C_{i,j}$  le temps entre la boulangerie  $i \in \{1, \dots, 6\}$  et le café  $j \in \{1, \dots, 6\}$ . Par exemple, on a  $C_{3,4} = 10$ , ce qui signifie qu'il y a dix minutes de trajet entre la boulangerie numéro 3 et le café numéro 4.

Afin de satisfaire la contrainte d'approvisionnement (que l'on appelle aussi la conservation de la masse), il faut que chaque boulangerie soit connectée à un et un seul café. Comme il y a le même nombre de boulangeries que de cafés, ceci implique que chaque café est également connecté à une

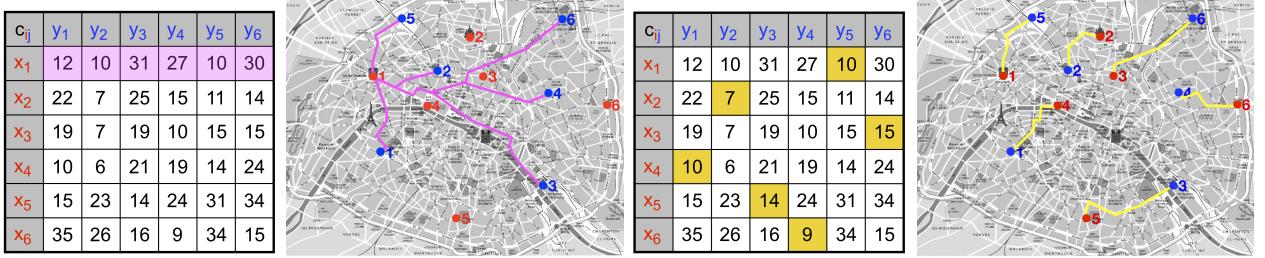


FIGURE 1 – Matrice de coût et connexions associées. Gauche : une ligne de la matrice coût. Droite : un exemple particulier de permutation.

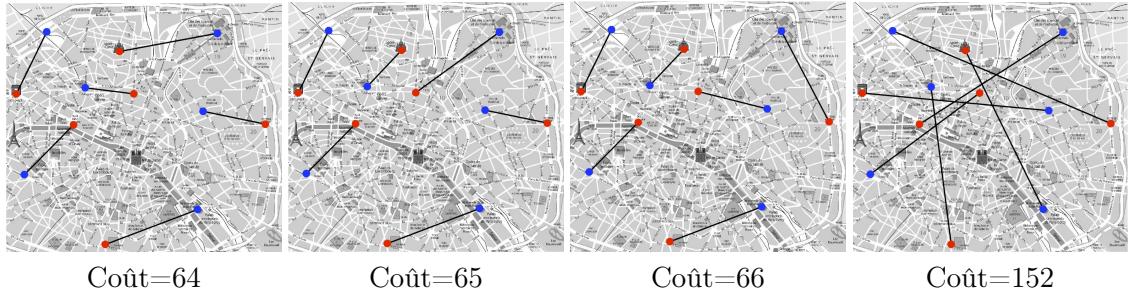


FIGURE 2 – Exemples de permutations avec différent coûts.

et une seule boulangerie. On va noter

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

un tel choix de connexions. Les deux images de droite de la figure 1 illustrent l'exemple

$$\sigma(1) = 5, \sigma(2) = 2, \sigma(3) = 6, \sigma(4) = 1, \sigma(5) = 3, \sigma(6) = 4. \quad (1)$$

La contrainte de conservation de masse signifie que  $\sigma$  est une bijection de l'ensemble  $\{1, \dots, 6\}$  dans lui-même. On dit aussi que  $\sigma$  est une permutation.

Le coût de transport associé à une telle bijection est la somme des coûts  $C_{i,\sigma(i)}$  sélectionnés par la permutation  $\sigma$ , c'est-à-dire

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

Par exemple, pour la bijection (1) montrée à la figure 1, on obtient comme coût

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

Le problème de Monge consiste à chercher une permutation  $\sigma$  qui a le coût minimum, c'est-à-dire résoudre le problème d'optimisation

$$\min_{\sigma \in \Sigma_6} \text{Coût}(\sigma), \quad (3)$$

où l'on a noté  $\Sigma_6$  l'ensemble des permutations de l'ensemble  $\{1, \dots, 6\}$ .

La figure 2 montre que la permutation (1) n'est pas la meilleure : il existe par exemple une autre permutation qui a un coût de 64. Mais est-ce la meilleure ? Il se trouve que oui, on peut en

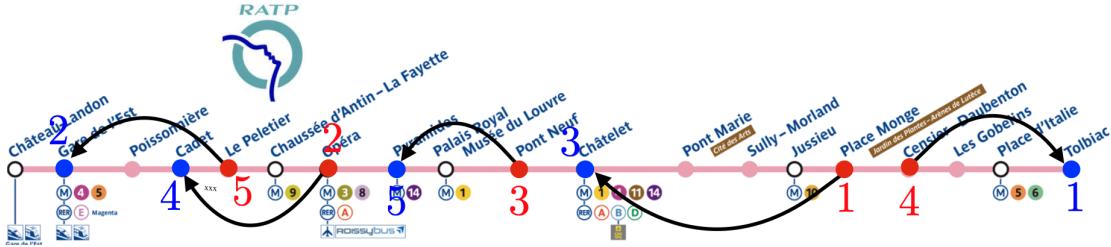


FIGURE 3 – Le transport optimal en 1D le long d'une ligne de métro. La bijection optimale est  $\sigma : (1, 2, 3, 4, 5) \mapsto (3, 4, 5, 1, 2)$ .

effet tester sur un ordinateur toutes les permutations de  $\{1, \dots, 6\}$  et calculer leur coût. Combien y a-t-il de permutations au total ? Pour effectuer ce dénombrement, on voit qu'il y a six choix d'affectation possible de 1 à  $\sigma(1) \in \{1, \dots, 6\}$ , puis cinq choix possibles pour affecter 2 à  $\sigma(2) \in \{1, \dots, 6\} - \{\sigma(1)\}$ , et ainsi de suite. Le nombre total de possibilités est donc  $6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$  que l'on note  $6!$ . Si l'on considère un nombre  $n$  de boulangeries, alors le nombre de permutations à tester pour trouver la meilleure est  $n! = n \times (n - 1) \times \dots \times 2 \times 1$ . Ce nombre croît extrêmement vite avec  $n$ , par exemple  $70! \approx 1, 198 \times 10^{100}$ , à comparer avec les  $10^{11}$  neurones dans le cerveau et les  $10^{79}$  atomes dans l'univers. Cette stratégie de recherche exhaustive n'est donc possible que pour de toute petites valeurs de  $n$ .

## 1.2 En 1D et 2D

Il aura fallu près de 200 ans pour que des idées nouvelles émergent pour calculer efficacement un transport optimal  $\sigma$  même pour des grandes valeurs de  $n$ . Avant d'expliquer ces avancées mathématiques, commençons par un cas dans lequel le transport optimal se calcule facilement. Le cas le plus élémentaire est lorsque les points à appairer sont le long d'un axe 1D, par exemple si les cafés et les boulangeries sont situés le long d'une ligne de métro. Il faut également que le coût  $C_{i,j}$  soit la distance le long de cet axe (par exemple le temps de trajet en métro entre les stations). On se place à gauche de tous les points en jeu et on parcourt la ligne de métro de gauche à droite. Le premier point rouge est associé avec le premier point bleu, le deuxième point rouge avec le deuxième point bleu, etc. Ce procédé est illustré à la figure 3. Le temps de calcul nécessaire pour calculer le transport optimal en métro est donc le temps nécessaire pour classer les indices. L'algorithme le plus simple pour effectuer un classement est celui utilisé habituellement pour trier un jeu de  $n$  cartes : il s'agit du tri par insertion, qui insère itérativement chaque carte à sa place par rapport aux cartes déjà classées. Il effectue  $n(n - 1)/2$  comparaisons. Pour  $n = 70$ , ceci nécessite donc seulement 2415 opérations, ce qui rend la méthode utilisable, au contraire de la recherche exhaustive de toutes les  $n!$  permutations. On dispose d'algorithmes encore plus rapides (par exemple le tri fusion), qui effectuent de l'ordre de  $n \log(n)$  opérations, et donc pour  $n = 70$ , de telles méthodes nécessitent moins de 1000 opérations.

Malheureusement, il n'est plus possible d'utiliser cette technique de classement dans des cas plus généraux. Pour des points en dimension 2, si on prend comme coût  $C_{i,\sigma(i)}$  la distance euclidienne (la distance en vol d'oiseau) entre les points, alors Gaspard Monge a montré dans son papier original (voir la figure 4, à gauche) qu'un transport optimal ne peut pas contenir de croisements. Par exemple, comme le montre la figure 4 (à droite), si l'on trace tous les segments entre les points  $i \mapsto j = \sigma(i)$

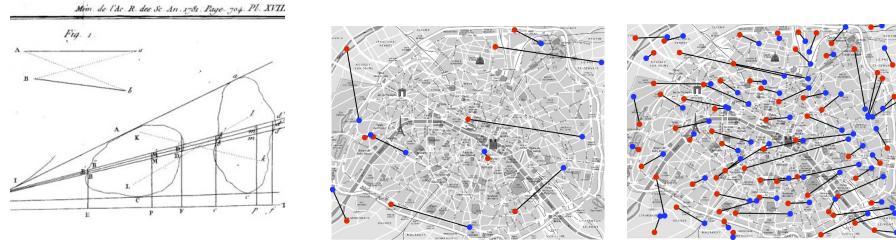


FIGURE 4 – Gauche : extrait de l'article de Monge [7]. Droite : le transport optimal en 2D pour un coût euclidien.

que l'on relie par la bijection définie par un  $\sigma$  optimal, ceux-ci ne se croisent jamais.

Cette observation géométrique n'est cependant pas suffisante pour calculer un transport optimal en 2D : il existe en effet beaucoup de permutations  $\sigma$  telles que les segments associés ne se croisent pas. Il va falloir analyser de façon plus fine la structure des permutations optimales afin de pouvoir les calculer de façon efficace. Nous allons maintenant voir comment Leonid Kantorovitch a reformulé le problème de Monge afin d'y parvenir.

## 2 Le Transport Optimal de Kantorovitch

Leonid Kantorovitch est un mathématicien et économiste soviétique qui a révolutionné la théorie du transport optimal pendant les années 40. Ses recherches sont issues de considérations pratiques qui l'ont occupé avant et après la seconde guerre mondiale. Il y a joué un rôle important pour assurer une distribution optimale des ressources, en particulier durant le siège de Léningrad. Il a par la même occasion participé au développement de l'optimisation moderne, laquelle a eu un impact énorme dans de très nombreux domaines appliqués. Il a ainsi obtenu en 1975 le prix Nobel d'économie, car les premières applications (mais certainement pas les seules !) de sa théorie se sont manifestées dans ce domaine.

### 2.1 Le problème de Kantorovitch

L'idée centrale de Kantorovitch [6] est de modifier le problème de Monge en remplaçant l'ensemble des permutations par un ensemble plus grand mais plus simple. Tout d'abord on remarque que l'on peut représenter une permutation  $\sigma \in \Sigma_n$  à l'aide d'une matrice de permutation  $P$  qui est une matrice binaire (remplie de 0 et de 1) de taille  $n \times n$  telle que  $P_{i,j} = 0$  sauf si  $j = \sigma(i)$  auquel cas  $P_{i,\sigma(i)} = 1$ . Par exemple, pour  $n = 3$  points, les permutations  $(1, 2, 3) \mapsto (1, 2, 3)$  (l'identité),  $(1, 2, 3) \mapsto (3, 2, 1)$  et  $(1, 2, 3) \mapsto (2, 1, 3)$  sont respectivement représentées par les matrices de taille  $3 \times 3$

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

Dans la suite, on note  $\mathcal{P}_n$  l'ensemble des  $n!$  matrices de permutation de taille  $n \times n$ .

Comme la matrice est binaire, avec seulement  $n$  éléments non nuls égaux à 1, on peut remplacer la somme de  $n$  termes qui apparaît dans  $\text{Coût}(\sigma)$  défini en (2) par une somme sur l'ensemble des

$n \times n$  indices  $(\textcolor{red}{i}, \textcolor{blue}{j})$ , c'est-à-dire que si  $P$  est la matrice de permutation associée à  $\sigma$ , on a

$$\text{Coût}(\sigma) = \sum_{i=1}^n \sum_{j=1}^n P_{\textcolor{red}{i}, \textcolor{blue}{j}} C_{\textcolor{red}{i}, \textcolor{blue}{j}}.$$

On peut ainsi remplacer le problème de Monge (3) par le problème équivalent

$$\min_{P \in \mathcal{P}_n} \sum_{i=1}^n \sum_{j=1}^n P_{\textcolor{red}{i}, \textcolor{blue}{j}} C_{\textcolor{red}{i}, \textcolor{blue}{j}}. \quad (4)$$

Le génie de Kantorovitch a été de remarquer que l'on peut remplacer l'ensemble discret  $\mathcal{P}_n$  (c'est-à-dire composé d'un ensemble fini, mais très grand, de  $n!$  matrices) par un ensemble « continu » (donc en particulier infini) mais plus simple. On remarque en effet que les matrices de permutation de  $\mathcal{P}_n$  sont exactement les matrices qui ont un et un seul 1 le long de chaque ligne et de chaque colonne. Ceci peut aussi s'exprimer comme le fait qu'une matrice de permutation est une matrice binaire dont la somme de chaque ligne et de chaque colonne vaut 1, c'est-à-dire

$$\mathcal{P}_n = \left\{ P \in \{0, 1\}^{n \times n} \mid \forall \textcolor{red}{i}, \sum_{\textcolor{blue}{j}} P_{\textcolor{red}{i}, \textcolor{blue}{j}} = 1, \forall \textcolor{blue}{j}, \sum_{\textcolor{red}{i}} P_{\textcolor{red}{i}, \textcolor{blue}{j}} = 1 \right\}.$$

Ce qui rend cet ensemble très compliqué, c'est la contrainte binaire, c'est-à-dire que ces matrices sont contraintes à être dans  $\{0, 1\}^{n \times n}$ . Kantorovitch propose alors de « relaxer » cette contrainte en supposant simplement que les entrées de  $P$  sont entre 0 et 1. Ceci définit un ensemble plus grand, l'ensemble des matrices bistrochastiques

$$\mathcal{B}_n \stackrel{\text{def.}}{=} \left\{ P \in [0, 1]^{n \times n} \mid \forall \textcolor{red}{i}, \sum_{\textcolor{blue}{j}} P_{\textcolor{red}{i}, \textcolor{blue}{j}} = 1, \forall \textcolor{blue}{j}, \sum_{\textcolor{red}{i}} P_{\textcolor{red}{i}, \textcolor{blue}{j}} = 1 \right\}. \quad (5)$$

Le problème de Kantorovitch s'obtient en effectuant ce remplacement dans (4), afin de résoudre

$$\min_{P \in \mathcal{B}_n} \sum_{i=1}^n \sum_{j=1}^n P_{\textcolor{red}{i}, \textcolor{blue}{j}} C_{\textcolor{red}{i}, \textcolor{blue}{j}}. \quad (6)$$

L'immense avantage du problème de Kantorovitch (6) par rapport à celui de Monge (4) est que l'ensemble des matrices bistrochastiques est convexe, c'est-à-dire que si l'on considère deux matrices bistrochastiques  $P, Q \in \mathcal{B}_n$ , alors leur moyenne  $\frac{P+Q}{2} \in \mathcal{B}_n$  est encore bistrochastique. Ceci n'est pas vrai pour les matrices de permutation, puisque la moyenne de deux matrices binaires  $(P, Q)$  n'est pas binaire (sauf bien sûr si  $P = Q$ ). Cette convexité est la clef pour le développement d'algorithmes efficaces. Cette nouvelle formulation a en effet pu bénéficier d'une deuxième révolution initiée par George Dantzig [4], qui, à la même époque, a proposé l'algorithme du simplexe. Celui-ci permet de résoudre efficacement une certaine classe de problèmes d'optimisation convexe : les problèmes de programmation linéaire, dont (6) est un cas particulier. Dans le cas du problème de Kantorovitch, il existe en effet un algorithme du simplexe qui a une complexité de l'ordre de  $n^3$  opérations, ce qui permet de faire des calculs pour de grands  $n$ , de l'ordre de plusieurs milliers.

## 2.2 L'équivalence Monge–Kantorovitch

L'ensemble des matrices bistochastiques est plus grand que celui des matrices de permutations,  $\mathcal{P}_n \subset \mathcal{B}_n$ , de sorte que l'on a l'inégalité

$$\min_{P \in \mathcal{B}_n} \sum_{\textcolor{red}{i}=1}^n \sum_{\textcolor{blue}{j}=1}^n P_{\textcolor{red}{i},\textcolor{blue}{j}} C_{\textcolor{red}{i},\textcolor{blue}{j}} \leq \min_{P \in \mathcal{P}_n} \sum_{\textcolor{red}{i}=1}^n \sum_{\textcolor{blue}{j}=1}^n P_{\textcolor{red}{i},\textcolor{blue}{j}} C_{\textcolor{red}{i},\textcolor{blue}{j}} \quad (7)$$

entre les problèmes de Kantorovitch et de Monge. Mais de façon à première vue surprenante, un théorème fondamental dû à George Birkhoff et à John von Neumann [2, 11] assure qu'en fait il y a égalité entre les valeurs de ces deux minimisations. En effet, ce théorème montre qu'il existe toujours une matrice solution du problème de Kantorovitch qui est une matrice de permutation, de sorte qu'elle est aussi solution du problème de Monge. Attention cependant, en général il n'y a pas unicité des solutions de ces problèmes : il peut exister une matrice bistochastique solution du problème de Kantorovitch qui n'est pas une permutation. La conjonction de deux avancées spectaculaires, dues à Kantorovitch et à Dantzig, a permis de rendre le transport optimal applicable à des problèmes de grande taille, puisque l'algorithme du simplexe permet de résoudre en pratique ces problèmes.

## 2.3 Le cas pondéré

Outre son intérêt pratique, la formulation de Kantorovitch a aussi permis de généraliser le problème initial de Monge, en donnant le bon cadre pour le formaliser et l'étudier mathématiquement. En effet, le problème de Monge est très limité. Que se passe-t-il par exemple s'il n'y a pas le même nombre  $n$  de cafés et  $m$  de boulangeries ? Le problème initial (3) n'a pas de solution, car on ne peut pas mettre en bijection deux ensembles de tailles différentes. Le bon concept n'est pas le nombre de boulangeries et de cafés, mais plutôt les distributions  $(\textcolor{red}{a}_1, \dots, \textcolor{red}{a}_n)$  de production (associées aux boulangeries) et les distributions  $(\textcolor{blue}{b}_1, \dots, \textcolor{blue}{b}_m)$  de consommation des cafés. Par exemple, si la première boulangerie produit 45 croissants par jour, on prendra  $\textcolor{red}{a}_1 = 45$ , de même  $\textcolor{blue}{b}_3 = 34$  signifie que le 3<sup>e</sup> café consomme 34 croissants par jour. Dans le cas initialement considéré, où  $n = m$ , toutes les quantités  $\textcolor{red}{a}_i$  et  $\textcolor{blue}{b}_j$  sont égales à 1. Mais dans de nombreux cas concrets, ces quantités sont quelconques. Ces quantités doivent être positives, et vérifier

$$\textcolor{red}{a}_1 + \dots + \textcolor{red}{a}_n = \textcolor{blue}{b}_1 + \dots + \textcolor{blue}{b}_m,$$

de sorte qu'il y ait autant de production que de consommation. La construction de Kantorovitch s'adapte naturellement à ce cas de distributions générales, en remplaçant les matrices bistochastiques (5) par des matrices de « couplage » qui satisfont la contrainte de conservation de la masse

$$\mathcal{B}(\textcolor{red}{a}, \textcolor{blue}{b}) \stackrel{\text{def.}}{=} \left\{ P \in \mathbb{R}_+^{n \times m} \mid \forall \textcolor{red}{i}, \sum_{\textcolor{blue}{j}} P_{\textcolor{red}{i},\textcolor{blue}{j}} = \textcolor{red}{a}_i, \forall \textcolor{blue}{j}, \sum_{\textcolor{red}{i}} P_{\textcolor{red}{i},\textcolor{blue}{j}} = \textcolor{blue}{b}_j \right\}.$$

Dans le cas initial où  $n = m$  et  $\textcolor{red}{a}_i = \textcolor{blue}{b}_j = 1$ , alors  $\mathcal{B}(\textcolor{red}{a}, \textcolor{blue}{b}) = \mathcal{B}_n$  et l'on retrouve des matrices bistochastiques. Dans le cas général, à chaque fois qu'une entrée  $P_{\textcolor{red}{i},\textcolor{blue}{j}}$  est non-nulle, ceci signifie que l'on transfère de la « masse » (ici une certaine quantité de croissants) entre  $\textcolor{red}{i}$  et  $\textcolor{blue}{j}$ . Comme le montre la figure 5, on peut visualiser de différentes façons une telle matrice  $P$  couplant deux distributions  $(\textcolor{red}{a}, \textcolor{blue}{b})$ . Contrairement au cas des matrices bistochastiques, pour lequel il y a toujours une solution

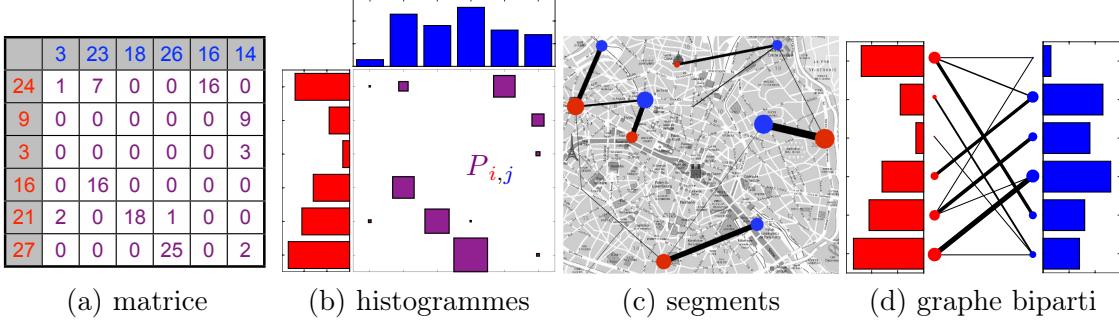


FIGURE 5 – Différentes façons de représenter une matrice de couplage  $P \in \mathcal{B}(a, b)$  : (a) un tableau de nombres dont les lignes et colonnes ont des sommes prescrites ; (b) un histogramme bidimensionnel dont la taille de carré est proportionnelle à  $P_{i,j}$  ; (c) un ensemble de segments dont la largeur est proportionnelle à  $P_{i,j}$ . (d) un graphe biparti, c'est-à-dire avec deux groupes de points reliés par des arêtes.

qui est une permutation, ici un couplage optimal  $\mathcal{B}(a, b)$  peut avoir plus d'une seule entrée non-nulle  $P_{i,j}$  le long d'une ligne indexée par  $i$  (voir la figure 5). Ceci signifie que cette boulangerie  $i$  est connectée à plusieurs cafés, de sorte que sa production est alors séparée en plusieurs lots de croissants distribués, tout en satisfaisant la contrainte de conservation de la masse  $\sum_j P_{i,j} = a_i$ .

Le problème de Kantorovitch qui généralise (6) s'écrit alors

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

ce qui signifie que l'on doit payer un coût  $C_{i,j}$  à chaque fois que l'on transfert une unité de masse entre  $i$  et  $j$ . Tout comme le problème original (6), on peut le résoudre de façon efficace avec l'algorithme du simplexe. La figure 5 montre un exemple de couplage optimal.

### 3 Les applications

Bien que les motivations initiales de Monge et Kantorovitch aient été respectivement militaires et économiques, le transport optimal a trouvé d'innombrables applications, à la fois théoriques et concrètes. Sur le plan mathématique, on peut considérer des distributions « continues » de masses, en quelque sorte la limite quand le nombre de points  $n$  tend vers l'infini. Ceci permet de définir le problème de transport entre des mesures de probabilités quelconques. Ce point de vue théorique est extrêmement fructueux, et c'est le mathématicien français Yann Brenier qui a le premier montré l'équivalence dans le cadre continu des formulations de Monge et de Kantorovitch [3]. Ces travaux pionniers ont montré la connexion entre le problème de transport et les équations aux dérivées partielles, et ont débouché, entre autres, sur les médailles Fields de Cédric Villani (2010) et Alessio Figalli (2018).

Le transport optimal est depuis peu au cœur de problématiques plus appliquées en sciences des données, en particulier pour résoudre des problèmes en traitement d'image et en apprentissage machine. La première idée, la plus immédiate, est d'utiliser la bijection  $\sigma$  afin de transformer des données, par exemple des images. Dans ce cas, on considère les pixels  $(x_i)_{i=1}^n$  et  $(y_j)_{j=1}^m$  de deux images couleur. Chaque pixel  $x_i, y_j \in \mathbb{R}^3$  est un vecteur de dimension 3, qui représente les intensités

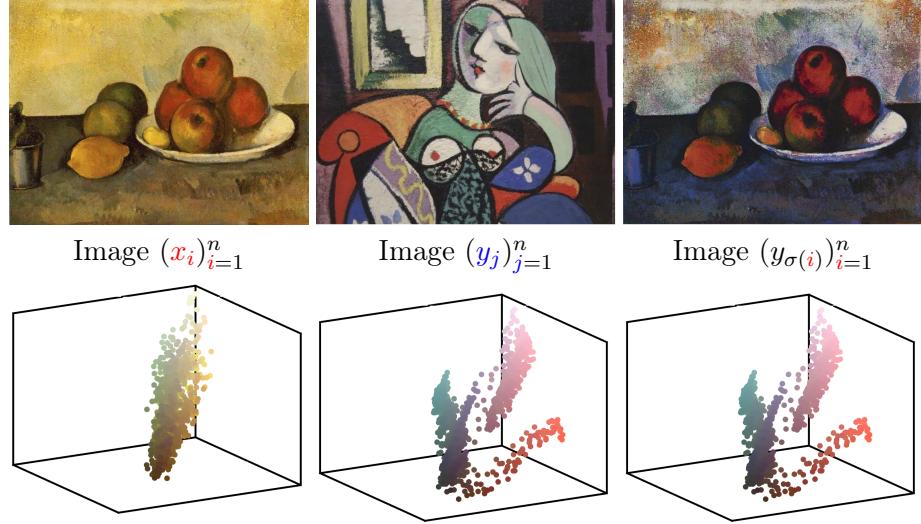


FIGURE 6 – Exemple de transfert de palettes de couleurs à l'aide du transport optimal. Haut : les pixels sont sur la grille d'affichage pour former une image couleur. Bas : les pixels sont placés à leurs positions dans  $\mathbb{R}^3$  pour former un nuage de points.

de chacune des trois couleurs élémentaires, rouge, vert et bleu. Afin de changer les couleurs de la première image, et lui imposer la palette de la deuxième image, on calcule le transport  $\sigma$  pour la matrice de coût  $C_{i,j} = \|x_i - y_j\|^2$  (c'est-à-dire le carré de la norme euclidienne dans  $\mathbb{R}^3$ ), c'est-à-dire le carré de la distance euclidienne entre les pixels. L'image avec les couleurs modifiées est  $(y_{\sigma(i)})_{i=1}^n$ , c'est à dire que l'on remplace dans la première image le pixel  $x_i$  par le pixel  $y_{\sigma(i)}$ . Cette image ressemble à la première, mais a la palette de couleurs de la deuxième image. La figure 6 illustre ce procédé pour imposer la palette de couleurs de Picasso à un tableau de Cézanne.

On peut également utiliser le transport optimal pour des problèmes plus difficiles, en n'utilisant que de façon indirecte la bijection  $\sigma$  ou bien la matrice de couplage optimal  $P \in \mathcal{B}(a, b)$ . L'idée centrale est que la quantité associée à un couplage optimal  $P$  solution de (8)

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

définit en quelque sorte l'effort nécessaire pour déplacer la masse de la distribution  $a$  vers la distribution  $b$ . Elle permet donc de quantifier combien ces deux distributions sont « proches ». Par exemple, si  $C_{i,j} = \|x_i - y_j\|^2$  est le carré de la distance euclidienne entre des points, alors la quantité  $W(a, b)^{1/2}$  est une distance entre les distributions, en particulier elle vérifie  $W(a, b) = 0$  si et seulement si  $a = b$ , et elle vérifie l'inégalité triangulaire. Ces propriétés sont très importantes pour permettre d'appliquer le transport à des problèmes pratiques.

Un exemple typique d'application de cette quantité  $W$  consiste à calculer des barycentres entre des distributions [1]. La figure 7 montre un exemple où l'on considère trois distributions  $a, b, c$  (montrées aux trois sommets du triangle) qui sont des distributions uniformes de masse à l'intérieur de formes 3D (c'est-à-dire que la masse  $a_i$  associée au  $i^{\text{e}}$  point est 0 à l'extérieur de la première forme et prend une valeur constante à l'intérieur). On calcule un barycentre pondéré de ces trois distributions en imitant le fait que dans un espace Euclidien, le barycentre pondéré  $r$  de trois points

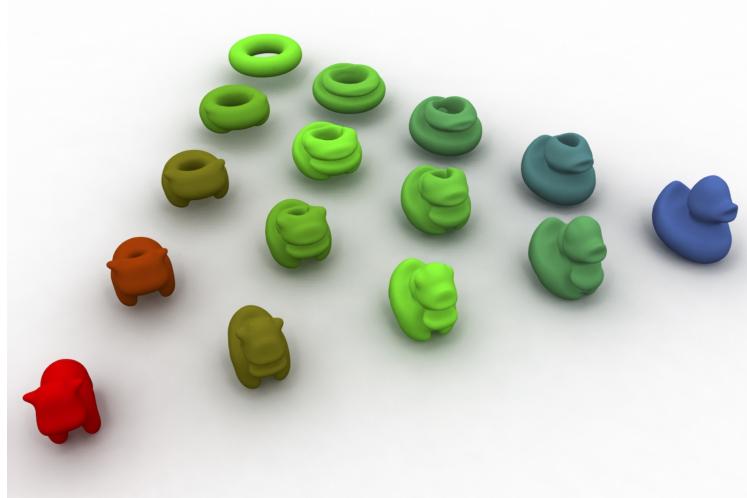


FIGURE 7 – Exemple d'interpolation barycentrique entre des formes 3D, obtenu en minimisant (9).



FIGURE 8 – Exemples d'histogrammes de distributions des mots, qui peuvent être utilisés pour comparer les discours d'Obama et de Lincoln (seuls les mots les plus fréquents sont montrés).

$x, y, z$  minimise la somme des distances au carré

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

où les poids  $(\alpha, \beta, \gamma)$  sont les pondérations du barycentre, qui sont des réels positifs et tels que  $\alpha + \beta + \gamma = 1$ . Le barycentre pondéré  $d$  de  $(a, b, c)$  minimise ainsi la somme pondérée de distances de transport optimal

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

En modifiant les poids  $(\alpha, \beta, \gamma)$ , on modifie la forme obtenue en se déplaçant à l'intérieur d'un triangle de transport optimal. On peut utiliser cette distance  $W$  pour bien d'autres applications où l'on doit comparer des distributions de probabilité. C'est le cas en apprentissage machine, par exemple pour comparer des textes à l'aide des distributions des mots qui les composent. La figure 8 illustre les histogrammes d'apparition des mots pour deux textes, où la taille des lettres du mot  $i$  est proportionnelle à la masse  $a_i$ . Une question difficile dans ce cas est de savoir quelle matrice de coût  $C_{i,j}$  utiliser entre deux mots  $(i, j)$ . Il s'agit d'un travail de linguistique (caractériser la proximité sémantique entre des mots du langages), que l'on peut chercher à résoudre en même temps que le transport optimal [5].

## Conclusions

Le transport optimal a connu de nombreuses révolutions. Sous l’impulsion de mathématiciens tels que Monge, Kantorovitch, Dantzig et Brenier, il est progressivement devenu un outil théorique et numérique fondamental. Il est maintenant au cœur de questions importantes en science des données pour modéliser, résoudre numériquement et analyser théoriquement les problèmes de l’apprentissage machine. Les opportunités pour développer de nouvelles théories et des algorithmes performants sont immenses. Pour plus d’informations sur les aspects théoriques du transport optimal, on pourra consulter les livres [10, 9]. Les aspects numériques et applicatifs sont couverts dans le livre [8].

## Remerciements

Je tiens à remercier Vincent Beck, Gwenn Guichaoua et Marie-Noëlle Peyré pour leurs relectures attentives.

## Références

- [1] Martial Aguech 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] 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. *to appear in Fundation and Trends in Machine Learning*, 2018.
- [9] Filippo Santambrogio. *Optimal transport for applied mathematicians*. Birkhauser, 2015.
- [10] Cedric Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics Series. American Mathematical Society, 2003.
- [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.