Introduction au Transport Optimal et Probabilités

Matthieu Boyer

Cours TalENS 2 2025-2026

Table des matières

Ce polycopié ne doit pas être vu comme un remplacement au cours, mais simplement comme un résumé du contenu qui permet d'être sûr de ne rien manquer.

1 Problème de Monge

Gaspard Monge, ingénieur militaire, révolutionnaire et fondateur de l'École Polytechnique et de l'École Normale Supérieure (entre autres), s'est posé le problème suivant dans son *Mémoire sur la théorie des déblais et des remblais* :

> Deux volumes équivalents étant donnés, les décomposer en particules infiniment petites se correspondant deux à deux, de telle façon que la somme des produits des chemins parcourus en transportant chaque parcelle sur celle qui lui correspond, par le volume de la parcelle transportée, soit un minimum.
>
> *Gaspard Monge*

Autrement dit :

Étant donné des sacs de sable dans n camps, quelle est la manière la moins fatigante de construire n murs à des endroits différents, sachant que les endroits sont à des distances plus ou moins grandes de chaque camp ?

Ce problème fait partie de la grande classe des problèmes d'optimisation, et est fondamental en théorie des probabilités, puisqu'il est à la base de la théorie du transport optimal qui donne des manières de trouver des distributions de probabilités "moyennes".

1.1 Bases de probabilités

On ne rentrera pas ici dans les "vraies" bases de la théorie de la mesure, et notamment sur la notion de tribu.

Définition 1.1 Une mesure sur un espace $\mathcal{X}$ muni d'une tribu ensemble Σ de parties de X est une application $\mu : \Sigma \to \mathbb{R}$ telle que :

- $\mu(\varnothing) = 0$
- $\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)$
- $\mu\left(\bigcup_{i=0}^{\infty} A_i\right) = \sum_{i=0}^{\infty} \mu(A_i)$ si les A_i sont deux à deux disjoints
- $A \subseteq B \Rightarrow \mu(A) \leq \mu(B)$.

On dit que $\mu(\mathcal{X})$ est la masse totale de μ.

Intuitivement, une mesure est une application qui donne la quantité d'éléments dans une partie d'un espace.

Dans notre cas, on s'intéressera principalement aux cas $|\mathcal{X}| \hookrightarrow \mathbb{N}$ et $X = \mathbb{R}^d$.

Définition 1.2 Sur $\mathcal{X}$ au plus dénombrable, on a une mesure dite de comptage qui a chaque partie de X associe son nombre d'éléments.

Sur $\mathcal{X} = \mathbb{R}^d$ on a une mesure λ dite mesure de Lebesgue qui à chaque partie de X associe son volume. En particulier, $\lambda\left(\prod_i [a_i, b_i]\right) = \prod_i |b_i - a_i|$.

Définition 1.3 Une mesure de probabilité est une mesure positive de masse 1. Une mesure μ a densité par rapport à la mesure de Lebesgue s'il existe une fonction p telle que

$$\mu(A) = \int_A p(x)\,\mathrm{d}\lambda(x)$$

Le symbole intégrale $\int$ signifie ici que pour tout point x dans A, on construit un petit pavé autour de A, qu'on calcule son volume $\mathrm{d}\lambda(x)$ et qu'on le multiplie par la densité de μ en x $p(x)$. Autrement dit, dans le volume $\mathrm{d}\lambda(x)$ autour de x, il y a $p(x)$ particules. On trouve donc le nombre total de particules dans A en sommant les nombres de particules autour de tout point x de A.

Plus généralement, il faut entendre

$$\int_A f(x)\,\mathrm{d}\alpha(x)$$

l'intégrale de f sur A par rapport à α comme le produit de la valeur de f en x par la "quantité" d'éléments autour de x définie par α. Cette intuition est en réalité une "définition" dans le cas réel. L'intégrale de f entre a et b selon la mesure de Lebesgue sur $\mathbb{R}$ (qui donc à $[a, b]$ associe la mesure $b - a$) permet de définir l'aire (algébrique) sous la courbe de f.

Vous verrez cette année que de plus, $\frac{\mathrm{d}}{\mathrm{d}x}(x \mapsto \int_0^x f(t)\,\mathrm{d}t) = f$ et que l'intégrale est une forme d'anti-dérivée sur $\mathbb{R}$. Il faut voir ce fait ainsi : la dérivée décrit la manière dont la valeur évolue, l'intégrale recouvre la courbe à partir d'un point et de l'évolution de celle-ci. C'est une forme d'inverse à la méthode de Newton de discrétisation que vous connaissez peut-être.

Définition 1.4 Pour $x \in \mathbb{R}^d$, on définit le dirac en x comme la mesure de probabilité δ_x qui vaut 1 sur $\{x\}$ et 0 ailleurs.

Pour $\mu \in \mathbb{R}^d, \Sigma \in \mathcal{M}_{d,d}(\mathbb{R})$, on définit la gaussienne centrée en μ de covariance Σ par sa densité $p(x) = \exp\left(-\frac{1}{2}\Sigma^{-1}\,(x-\mu)^2\right)$.

Définition 1.5 Une variable aléatoire X à valeurs dans E est une fonction dite mesurable de $\mathcal{X}$ dans E. La probabilité que X soit à valeurs dans $A \subseteq E$ est la mesure $\mu(X^{-1}(A))$.

Autrement dit, c'est la quantité (au sens de μ) d'antécédents des éléments de A dans par X.

On peut redéfinir ainsi l'espérance et la variance du variable aléatoire (et vous pouvez vérifier que c'est bien cohérent avec votre définition) :

Définition 1.6 L'espérance $\mathbb{E}[X]$ et la variance $\mathbb{V}[X]$ d'une variable aléatoire X à valeurs dans E de loi $\mathbb{P}(X = x) = p(x)$ (si elles existent) sont données par

$$\mathbb{E}[X] = \int_E x \, \mathrm{d}p(x) \quad \mathbb{V}[X] = \mathbb{E}\left[X - \mathbb{E}\left[X\right]\right]$$

Remarquez que quand X prend un ensemble fini de valeurs x_i, on retrouve $\mathbb{E}[X] = \sum_i x_i \mathbb{P}\left(X = x_i\right)$.

1.2 Formulation de Monge

On considère dans la suite deux mesures de probabilité $\alpha \in \mathcal{P}(\mathcal{X}), \beta \in \mathcal{P}(\mathcal{Y})$.

Définition 1.7 Pour $T : \mathcal{X} \to \mathcal{Y}$, on définit la mesure $T_\sharp \alpha$ par $T_\sharp \alpha(B) = \alpha(T^{-1}(B))$.

Autrement dit, $T_\sharp \alpha(B)$ est la quantité (selon α) d'antécédents des éléments de B par T.

La formulation de Monge du problème de transport optimal de α à β est la suivante :

Définition 1.8 Soit c une fonction de coût de $\mathcal{X}$ à $\mathcal{Y}$, c'est-à-dire une fonction positive de $\mathcal{X}$ dans $\mathcal{Y}$. Le problème de Monge associé à α, β, c est de calculer :

$$M = \inf_{T:\mathcal{X}\to\mathcal{Y}} \left\{ \int c(x, T(x)) \, \mathrm{d}\alpha(x) \,\middle|\, T_\sharp \alpha = \beta \right\} \tag{1}$$

M représente la manière optimale de déplacer tout le poids de α vers β, sachant le coût du déplacement d'un point de $\mathcal{X}$ vers un point de $\mathcal{Y}$.

Définition 1.9 Pour $c = \|\cdot\|$ une norme, on définit la p-distance de Wasserstein associée à c comme

$$W_p(\alpha, \beta) = \inf_{T_\sharp \alpha = \beta} \sqrt[p]{\int \|x - T(x)\|^p \, \mathrm{d}\alpha(x)}.$$

2 Calcul et Applications

2.1 Calcul discret

Dans le cas, plus simple, où $\alpha = \sum_{i=1}^N a_i \delta_{x_i}$ et $\beta = \sum_{j=1}^N b_j \delta_{y_j}$, les applications T telles que $T_\sharp \alpha = \beta$ correspondent à des assignations des x_i aux y_j, c'est à dire une permutation σ de l'ensemble $[\![1, N]\!]$ telle que $T(x_i) = y_{\sigma(i)}$.

Le problème d'optimisation qui correspond au transport optimal dans ce cas est le problème de couplage sur un graphe biparti complet :

Définition 2.1 Un graphe biparti complet est constitué d'un ensemble $V = A \sqcup B$ de *sommets* composé de deux parties A et B de tailles égales n. Pour chaque sommet de A on ajoute une *arête* à tout sommet de B avec des poids associés au coût entre les deux sommets. Un couplage sur un graphe biparti est le choix de n arêtes qui touchent tout sommet de A et tout sommet de B.

Dans l'Algorithme 1 on présente un pseudo-code pour l'algorithme hongrois qui permet de résoudre ce problème. L'idée de base étant de se ramener à chaque itération à un problème plus simple : une fois qu'un sommet est assigné,

Algorithme 1 Algorithme Hongrois pour le problème d'assignation bipartite

procedure MAXZ(C) ▷ Input : Cost matrix C of size $t \times t$
  $Z \leftarrow \{\}$
  while $0 \in C_{i,j}$, $\forall (i,j) \notin Z$ **do**
    $x \leftarrow$ row such that $C_{x,y}$ has the fewest marked 0 elements
    $y \leftarrow$ column such that $C_{x,y} = 0$ and C_y has the fewest marked 0 elements
    Add (x, y) to set Z
    Mark $C_{x,y}$ as an independent zero
return Set Z containing independent zeros
end procedure

procedure MINCOVER(C, Z) ▷ Input : Cost matrix C and set Z
  $coveredRows \leftarrow \{\}$ ▷ Rows covered with horizontal lines
  $coveredCols \leftarrow \{\}$ ▷ Columns covered with vertical lines
  while $C_{i,j} = 0$ and $i \notin coveredRows$ **do**
    if $0 \in C_i$ and $i \in Z$ **then**
      $coveredRows \leftarrow i$ ▷ Cover row i
    if $C_{i,j} = 0$ and $i \notin coveredRows$ **then**
      $coveredCols \leftarrow j$ ▷ Cover column j
return $coveredRows, coveredCols$
end procedure

procedure HUNGARIAN(C) ▷ Input : Cost matrix C of size $t \times t$
  for every row i and column j in C **do**
    $C_i \leftarrow C_i - \min(C_i)$ ▷ Subtract row minimum
    $C_j \leftarrow C_j - \min(C_j)$ ▷ Subtract column minimum
  $Z \leftarrow$ MAXZ(C)
  $cx, cy \leftarrow$ MINCOVER(C, Z)
  while $\text{len}(cx) + \text{len}(cy) < n$ **do**
    $minVal \leftarrow \min(C_{ij})$ such that $i \notin cx$ and $j \notin cy$
    for every row $i \notin cx$ and column $j \notin cy$ **do**
      $C_{ij} \leftarrow C_{ij} - minVal$
    $Z \leftarrow$ MAXZ(C)
    $cx, cy \leftarrow$ MINCOVER(C, Z)
return Set Z containing index (i, j) of assignments
end procedure

on le "supprime" virtuellement de la matrice de coût puis on recommence jusqu'à avoir terminé. Cet algorithme est dit *glouton* parce qu'il effectue à tout instant l'assignation la plus efficace.

Cet algorithme est assez efficace, pusiqu'il effectue un nombre d'opérations dit $\mathcal{O}(n^3)$ (lire « grand O de n^3 »), qui est de l'ordre de n^3 et plus mathématiquement, il existe une constante positive C telle que le nombre d'opérations $f(n)$ en fonction de n vérifie $f(n) \leq n^3$.

2.2 Applications du transport optimal

On revient au problème de l'analyse de données. On se donne donc une immersion (embedding) $\mathcal{X} \subseteq \mathbb{R}^d$ d'un ensemble de données.

Définition 2.2 Le barycentre de n points x_i pour une fonction d'énergie E et des poids λ_i est le point $\bar{x}$ tel que

$$\bar{x} = \text{argmin}_x \sum_{i=1}^{n} \lambda_i E(x, x_i).$$

Dans notre cas, on prendra $E = W_2^2$ la 2-distance de Wasserstein au carré, pour des questions de facilité du problème d'optimisation associé.

Canonisation On peut notamment chercher à savoir à quoi ressemble un exemple canonique de nos données. En particulier, dans le cas où les *features* (les d axes de l'espace d'immersion) sont disjointes, et représentent les résultats d'une expérience aléatoire répétée, en voyant chaque vecteur comme une distribution de probabilité, on peut trouver une distribution moyenne en calculant le barycentre des vecteurs. Calculer le barycentre de $\mathcal{X}$ nous permet donc de définir de manière efficace la distribution *moyenne* de notre ensemble de données, et donc, en revenant à travers l'immersion, un exemple canonique de notre ensemble de données de départ.

Réduction de dimension La distance de Wasserstein nous permet également de définir des points extrémaux. Notamment, on peut calculer des vecteurs de base a_k de sorte que chacun de nos points s'écrive comme le barycentre des a_k pour certains poids λ_k : $x_i = \text{argmin}_x \sum_{k=1}^{m} \lambda_k W_2^2(x, a_k)$. En prenant m petit devant n, ceci nous permet de réduire la dimension de notre espace de données. Ceci permet d'avoir un ensemble un peu plus général de données canoniques.

Architecture philosophique Finalement, quel est le rapport avec l'architecture ? Pour le trouver, il suffit de revenir à la définition originelle du problème de Monge. Comment déplacer des matériaux de manière à minimiser le coût de transport total ? Comment attribuer des courses à des taxis ? Ces deux problèmes semblent assez similaires, mais leur solution est similaire à la manière qu'on a de comparer les manières de représenter l'usage syntaxique de certaines déclinaisons dans différentes langues, ou de comparer les scans en 3D de différentes mains. Cédric Villani propose cette expérience, pour expliquer la beauté qu'il trouve au transport optimal (et qui lui a tout de même valu une médaille Fields en 2010, avant que son doctorant Alessio Figalli ne l'obtienne en 2018). Il relie ainsi deux domaines a priori éloignés des mathématiques (et même des sciences), la géométrie et la physique. Sans tomber dans un délire monomaniaque, il est aisé de voir en quoi obtenir des méthodes géométriques pour résoudre nombre de problèmes fondamentalement probabilistes ou obtenir des méthodes probabilistes pour résoudre nombre de problèmes fondamentalement géométriques. Le transport optimal permet de définir, selon la géométrie d'un espace donné, la géométrie de l'espace des probabilités sur cet espace et, par le transport des points, transporte la géométrie.

> L'expérience consiste à se donner une répartition de gaz dans l'espace, avec des fluctuations de densité d'une région à l'autre. On impose au gaz une nouvelle configuration, à atteindre en un temps limité, disons une minute. Le gaz obtempère, mais comme il est paresseux, il le fait en évoluant de manière à minimiser l'effort total (mesuré à chaque instant par l'énergie cinétique). Entre le temps initial et le temps final, on étudie les valeurs de l'entropie, qui mesure en un certain sens bien précis l'étalement du gaz (l'entropie est d'autant plus grande que la densité est faible). Si l'on vit dans un espace à courbure positive, alors la courbe d'entropie est concave ; en particulier elle est située au-dessus de la droite joignant les valeurs initiale et finale. La propriété de concavité est en fait caractéristique des espaces à courbure positive ; ce qui ouvre de nouveaux horizons pour étudier (voire pour redéfinir) les espaces à courbure positive. En mélangeant des notions d'ingénierie, de mécanique des fluides et de physique statistique, on a ainsi obtenu de nouveaux outils géométriques !
>
> *Cédric Villani*