RAPPORT: TRANSPORT OPTIMAL

ANDRIAMAHERISOA Liantsoa

Juin 2020

1 Introductions

Le principe du chemin le plus court guide la plupart des décisions dans les domaines de la vie et des sciences: lorsqu'un produit, une personne ou un seul élément d'information est disponible à un point donné et doit être accessible à envoyer à un point cible. Il convient de privilégier le moins d'effort possible. La théorie du transport optimal (TO) généralise cette intuition dans le cas où, au lieu de déplacer un seul élément à la fois, on se préoccupe du problème de déplacement simultanément de plusieurs éléments (ou une distribution continue de ceux-ci) d'un espace à un autre.
Introduite au 18ème siècle, ce n'est que récemment que cette théorie constitue un terrain fertile pour les chercheurs en informatique, en imagerie et plus généralement en sciences des données car cette théorie fournissait des outils très puissant pour étudier les distributions dans un contexte différent et plus abstrait, celui de comparer les distributions, facilement disponibles sous la forme d'un sac de caractéristiques ou d'informations. Maintenant que le TO s'est progressivement imposé comme un outil informatique, centré sur les applications à la science des données, notamment, l'imagerie et l'apprentissage automatique.

2 Fondements Théoriques

Ici, nous allons introduire les notions relatives d'appariements et de couplage optimaux entre vecteurs de probabilité (a, b), généralisant progressivement ce calcul pour un transporter des mesures discrètes (α, β), pour couvrir enfin le cadre général des mesures arbitraires.

2.1 Histogrammes et Mesures

On utilise de manière interchangeable les termes histogramme et vecteur de probabilité pour tout élément a $\in \sum_n$ qui appartient au simplexe de probabilité.

$$\Sigma_n \stackrel{\text{def.}}{=} \left\{ \mathbf{a} \in \mathbb{R}_+^n \ : \ \sum_{i=1}^n \mathbf{a}_i = 1 \right\}.$$

Pour des mesures discrètes avec des poids a et les emplacements $x_1, \ldots, x_n \in$ X. Noté:

$$\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i},$$

où σx est le Dirac à la position x, intuitivement une unité de masse est concentrée à l'infini au position x. Cette mesure décrit une mesure de probabilité si, en plus, a $\in \sum_n$ et plus généralement une mesure positive si tous les éléments du vecteur sont non négatifs.

Pour les mesures en générales: discrètes ou continues. On s'appuie sur l'ensemble des mesure de Randon M(X) sur l'espace X. X doit être accompagné d'une distance d, car on ne peut accéder à une mesure qu'en l'intégrant contre les fonctions continues, notés f∈C(X):

$$\int_{\mathcal{X}} f(x) \mathrm{d}\alpha(x) = \sum_{i=1}^n \mathbf{a}_i f(x_i).$$

2.2 Affectation et problème de Monge

Dans ce cas, on va avoir une matrice de coût

$$(\mathbf{C}_{i,j})_{i \in [\![n]\!], j \in [\![m]\!]}$$

En supposant n = m, le problème d'affectation optimal cherche une bijection σ dans l'ensemble Perm(n) des permutations d'éléments. On cherche alors à faire:

$$\min_{\sigma \in \mathrm{Perm}(n)} \frac{1}{n} \sum_{i=1}^n \mathbf{C}_{i,\sigma(i)}.$$

Une méthode naïve est d'évaluer cette fonction coût en utilisant toutes les permutations dans l'ensemble Perm(n). Cependant, cet ensemble a une taille n! Qui est énorme même pour des petits n. On va devoir utiliser des algorithmes efficaces pour optimiser le calcul. Le problème d'affection optimale peut avoir plusieurs solutions optimales, comme par exemple dans le cas de deux éléments identiques, on peut avoir deux affectations possibles qui sont tout les deux optimales.

Dans le cas de mesures discrètes, le problème Monge cherche une carte qui associe à chaque point x_i un seul point y_j et qui va pousser la masse de α vers la masse de β. Une telle carte T doit vérifier:

$$\forall j \in [\![m]\!], \quad \mathbf{b}_j = \sum_{i : T(x_i) = y_j} \mathbf{a}_i,$$

Dans le cas de mesures arbitraires (α, β), supposant deux espaces (X, Y) peuvent être liés par une carte T:X → Y qui minimise:

$$\min_T \left\{ \int_{\mathcal{X}} c(x, T(x)) \mathrm{d}\alpha(x) \ : \ T_\sharp \alpha = \beta \right\}.$$

La contrainte $\mathrm{T}_{\mp\alpha} = \beta$ veut dire que T pousse en avant la masse de α vers la masse de β, en utilisant la méthode de push-forward. La nature combinatoire du problème dans le cas discret et la complexité de la contrainte $\mathrm{T}_{\mp\alpha}$ qui est convexe font que le problème de Monge est très difficile.

2.3 La formulation relaxée de Kantorovich

Puisque le problème de Monge est trop ardu, il faut nous résigner à en résoudre un plus simple. C'est la stratégie proposée par Kantorovich. Plutôt que d'exiger que l'élément x d'une distribution α soit transporté de manière déterministe vers un point T(x) dans la distribution cible β, Kantorovich propose d'autoriser la répartition des éléments sur différents points dans la cible. Un plan de transport γ entre deux distributions discrètes α et β. La masse d'un point dans α peut être répartie sur plusieurs points dans β. γ(x,y) est une distribution de probabilités jointe qui indique la proportion d'élément au voisinage de x dans α que l'on va transporter au voisinage de y dans β. Pour que ce plan de transport soit compatible avec α et β il faut exiger les contraintes de marginalité . Cette flexibilité est alors codée en utilisant, à la place d'une permutation σ ou une carte T, une matrice de couplage $P \in \mathrm{R}(\mathrm{n} \times \mathrm{m})_+$, où P_i, j décrit la quantité de masse circulant de i vers j, ou de la masse trouvée en x_i vers y_j. Les couplages admissibles admettent une caractérisation beaucoup plus simple que les cartes Monge, où, on utilise la matrice vectorielle suivante:

$$\mathbf{P}\mathbb{1}_m = \left(\sum_j \mathbf{P}_{i,j} \right)_i \in \mathbb{R}^n \quad \text{and} \quad \mathbf{P}^{\mathrm{T}}\mathbb{1}_n = \left(\sum_i \mathbf{P}_{i,j} \right)_j \in \mathbb{R}^m.$$

L'ensemble des matrices U(a,b) est borné et défini par n+m contraintes d'égalité, et est donc un polytope convexe (la coque convexe d'un ensemble fini de matrices). De plus, alors que la formulation de Monge était intrinsèquement asymétrique, la formulation détendue de Kantorovich est toujours symétrique, dans le sens où un couplage P est dans U(a,b) si et seulement si P^T. est dans U(b,a), d'où:

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a},\mathbf{b}) \overset{\text{def.}}{=} \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{C},\mathbf{P}\rangle \overset{\text{def.}}{=} \sum_{i,j} \mathbf{C}_{i,j}\mathbf{P}_{i,j}.$$

C'est une problème linéaire, et comme c'est généralement le cas avec de telles problèmes, ses solutions optimales ne sont pas forcément uniques.

Matrices de permutation comme couplages: Pour une permutation σ ∈Perm(n), on écrit Pσ pour la matrice de permutation correspondantes:

$$\forall (i,j)\in[\![n]\!]^2,\quad (\mathbf{P}_\sigma)_{i,j} = \begin{cases} 1/n & \text{if} \quad j=\sigma_i, \\ 0 & \text{otherwise.} \end{cases}$$

Pour le cas des mesures discrètes, nous stockons dans la matrice C tous les coûts par paire entre les points dans les supports de α, β, à savoir $C_{i,j}$ = c(x_i, y_j), pour définir:

$$\mathcal{L}_c(\alpha,\beta) \overset{\text{def.}}{=} \mathrm{L}_{\mathbf{C}}(\mathbf{a},\mathbf{b}).$$

Par conséquent, la formulation de Kantorovich du TO entre les mesures discrètes est la même que le problème entre leurs vecteurs de probabilité de poids a,b sauf que la matrice de coût C dépend du support de α et β. Dans le cas des mesures arbitraires, L_c est étendu à des mesures arbitraires en considérant les couplages π∈M^1_+ (X×Y) qui sont des distributions conjointes sur l'espace produit. Le cas discret est une situation particulière où l'on impose que cette mesure de produit soit de la forme $\pi = \sum_{i,j} \mathrm{P}_{i,j}\delta(\mathrm{x}_i, y_j)$. Dans le cas général, la contrainte de conservation de la masse doit être réécrite en tant que contrainte marginale sur la distribution de probabilité conjointe.

$$\mathcal{U}(\alpha,\beta) \overset{\text{def.}}{=} \left\{\pi\in\mathcal{M}^1_+(\mathcal{X}\times\mathcal{Y}) \;:\; P_{\mathcal{X}\sharp}\pi=\alpha \quad\text{and}\quad P_{\mathcal{Y}\sharp}\pi=\beta\right\}.$$

2.4 Propriétés métriques du TO : Wasserstein

Pour définir une notion de distance entre deux distributions α et β, on utilise la distance de p-Wasserstein entre α et β. Pour la fonction de coût de transport C on choisit C(x,y) = D^p (x,y) où D(x,y) est une distance entre x et y et où p ≥ 1.

$$W_p(\alpha,\beta) \equiv [L_{D^p}(\alpha,\beta)]^{1/p} \quad \text{pour} \quad p\geq 1$$

On doit vérifier: W_p (α, γ) ≥ W^p (α, β) + Wp (β, γ). Dans certains cas spécifiques ont peut néanmoins W^p explicitement:

- distance entre deux distributions $\alpha = \delta_x$ et $\beta = \delta_y$, on vérifie immédiatement que W_p (δ_x, δ_y) = D(x,y)

- dans le cas p = 2, si l'on définit α' et β' comme les versions centrées de moyennes nulles α et β, qu'on définit m_α et m_β comme les moyennes respectives de α et β alors on a la décomposition $W_2(\alpha,\beta)^2 = W_2(\alpha',\beta')^2 + \|m_\alpha - m_\beta\|^2$.

- dans le cas p = 2 avec deux gaussiennes $\alpha = \mathrm{N}(m_\alpha, \sum_\alpha)$ et $\beta = \mathrm{N}(m_\beta, \sum_\beta$, on a: $W_2(\alpha,\beta)^2 = \|m_\alpha - m_\beta\|^2 + B(\sum_\alpha, \sum_\beta)$ où B est une métrique entre matrices de covariances que l'on sait calculer explicitement.

2.5 Problème Dual

Le problème de Kantorovich est un problème de minimisation convexe avec contraintes et, par conséquent, il peut naturellement être associé à un problème dual, qui est un problème de maximisation concave avec contraintes. Le problème de Kantorovich admet:

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a},\mathbf{b}) = \max_{(\mathbf{f},\mathbf{g})\in\mathbf{R}(\mathbf{C})} \langle \mathbf{f},\mathbf{a}\rangle + \langle \mathbf{g},\mathbf{b}\rangle,$$

où, l'ensemble des variables duals est:

$$\mathbf{R}(\mathbf{C}) \overset{\text{def.}}{=} \{(\mathbf{f},\mathbf{g})\in\mathbb{R}^n\times\mathbb{R}^m \;:\; \forall (i,j)\in[\![n]\!]\times[\![m]\!], \mathbf{f}\oplus\mathbf{g}\leq\mathbf{C}\}.$$

appelées potentiels de Kantorovich. Ce résulta est une conséquence directe du résultat plus général sur la forte dualité des programmes linéaires. Déduisons notre résultat en utilisant la dualité Langrangienne. D'où:

$$\min_{\mathbf{P}\geq 0} \max_{(\mathbf{f},\mathbf{g})\in\mathbb{R}^n\times\mathbb{R}^m} \langle \mathbf{C},\mathbf{P}\rangle + \langle \mathbf{a}-\mathbf{P}\mathbb{1}_m,\mathbf{f}\rangle + \langle \mathbf{b}-\mathbf{P}^{\mathrm{T}}\mathbb{1}_n,\mathbf{g}\rangle.$$

En échangeant le min et le max, on obtient:

$$\max_{(\mathbf{f},\mathbf{g})\in\mathbb{R}^n\times\mathbb{R}^m} \langle \mathbf{a},\mathbf{f}\rangle + \langle \mathbf{b},\mathbf{g}\rangle + \min_{\mathbf{P}\geq 0} \langle \mathbf{C}-\mathbf{f}\mathbb{1}_m{}^{\mathrm{T}} - \mathbb{1}_n\mathbf{g}^{\mathrm{T}},\mathbf{P}\rangle.$$

On en conclu grâce à ce remarque:

$$\min_{\mathbf{P}\geq 0} \langle \mathbf{Q},\mathbf{P}\rangle = \begin{cases} 0 & \text{if} \quad \mathbf{Q}\geq 0, \\ -\infty & \text{otherwise} \end{cases}$$

de sorte que la contrainte se lit

$$\mathbf{C}-\mathbf{f}\mathbb{1}_m{}^{\mathrm{T}} - \mathbb{1}_n\mathbf{g}^{\mathrm{T}} = \mathbf{C}-\mathbf{f}\oplus\mathbf{g}\geq 0.$$

Ainsi, la relation d'optimalité primale-dual pour le lagrangien permet de localiser le support du plan de transport optimal.