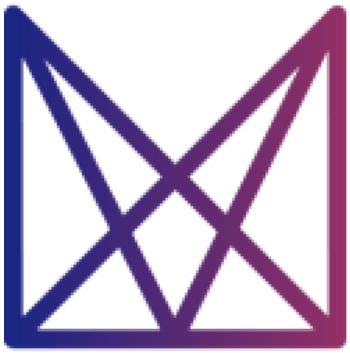

MATHÉMATIQUES
VISION
APPRENTISSAGE

Adaptation d'algorithmes pour des problématiques de transport optimal déséquilibré.

Mémoire du cours de Transport Optimal

Janvier 2022

Boammani Aser LOMPO, encadré par Gabriel PEYRE

Table des matières

1 Abstract **3**

2 Introduction **4**
2.1 Présentation du problème 4
2.2 Travaux précédents 4
2.3 Contribution de l'article 5
2.4 Notations 5

3 Présentation des résultats **7**
3.1 L'équation du Transport Optimal 7
3.2 Regularisation entropique 9
3.3 Algorithme 10
3.4 Expérience 12

4 Conclusion et LIEN avec le cours **12**

5 Bibliographie **14**

1 Abstract

L'objet de cet exposé est l'étude du problème de transport optimal dans un contexte où les distributions de départ et d'arrivée n'ont pas la même masse. En effet, la formulation de Kantorovich du problème du Transport Optimal est la suivante. Supposons que nous avons deux mesures de probabilité α et β définies sur deux ensembles X et Y et que supposons que nous disposons d'une fonction de coût $c(x, y)$ définie sur $X \times Y$. Le but est alors de trouver une mesure de probabilité γ sur $X \times Y$ qui minimise la quantité $\int_{X\times Y} c(x, y)\,\mathrm{d}P(x, y)$ et telle que $P^X_\sharp \gamma = \alpha$ et $P^Y_\sharp \gamma = \beta$. Ce problème tel que formulé ainsi ne peut pas être résolu si α et β ne sont plus des mesures de même masse. C'est dans ce contexte que va se mener l'étude qui va suivre. Nous montrerons une nouvelle formulation plus générale du problème du Transport Optimal capable de prendre en compte ce cas de figure, et nous proposerons une généralisation de l'algorithme de Sinkhorn pour le résoudre. Aussi nous montrerons comment cet algorithme s'applique au problème de barycentre de Wasserstein. Enfin nous montrerons quelques illustrations de la performance de l'algorithme sur des problèmes.

2 Introduction

2.1 Présentation du problème

La formulation de Kantorovich du problème du Transport Optimal est la suivante. Supposons que nous avons deux mesures de probabilité α et β définies sur deux ensembles X et Y et que supposons que nous disposons d'une fonction de coût $c(x, y)$ définie sur $X \times Y$. Le but est alors de trouver une mesure de probabilité γ sur $X \times Y$ qui minimise la quantité $\int_{X\times Y} c(x, y)\, \mathrm{d}P(x, y)$ et telle que $P^X_\sharp \gamma = \alpha$ et $P^Y_\sharp \gamma = \beta$. Une conséquence directe de cette formulation est que $\alpha(X) = \int_{x\in X} \mathrm{d}\alpha(x) = \int_{x\in X} \mathrm{d}P^X_\sharp \gamma(x) = \int_{X\times Y} \mathrm{d}\gamma = \gamma(X \times Y)$. De même $\beta(Y) = \gamma(X \times Y)$. Ce lien intrinsèque entre α, β et γ, interdit toute utilisation de cette formulation dans un contexte où $\alpha(X) \neq \beta(Y)$. Nous souhaitons nous libérer de cette contrainte afin de pouvoir adresser des problèmes plus généraux.

2.2 Travaux précédents

Le problème du Transport Optimal est vieux de plusieurs siècles. La première formulation de Monge :

$$\mathrm{argmin}_{\substack{T:X\longrightarrow Y \\ T_\sharp\alpha=\beta}} \left\{ \int_X c(x, T(x))\, \mathrm{d}\alpha(x) \right\}$$

a été résolue par Brenier qui publie le théorème suivant.

Théorème 1. Brenier

Supposons que $X = Y = \mathbb{R}^d$ et que $c(x, y) = ||x - y||^2$. Alors si α a une densité par rapport à la mesure de Lebesgue, il existe un unique couplage T qui minimise la quantité définie plus haut. De plus T est le gradient d'une fonction convexe ϕ.

La formulation de Kantorovich est venue apporter plus de souplesse dans la manière de transporter l'information de X à Y en permettant plus de connexions entre les éléments de X et ceux de Y au travers de la quantité $\gamma(x, y)$. Grâce aux travaux de Birkoff et de Von Neumann, il a été prouvé que dans le cas où X et Y sont des ensembles discrets de même cardinaux, les formulations de Kantorovich et de Monge sont équivalentes. Plus généralement, Brenier prouve aussi l'équivalence dans le cas où α a une densité et que c est le carré de la norme euclidienne.

La technique usuellement utilisée pour résoudre le problème de Kantorovich est celle de la régularisation entropique. Pour cela on introduit le problème de Schrödinger

$$\min_{\substack{\gamma\in\mathcal{M}_+(X\times Y) \\ P^X_\sharp\gamma=\alpha, P^Y_\sharp\gamma=\beta}} \int_{X\times Y} c(x, y)\, \mathrm{d}\gamma(x, y) + \epsilon \mathrm{KL}(\gamma|\alpha \otimes \beta)$$

où KL est utilisée comme fonction de régularisation pour obtenir une approximation γ_ϵ de la solution du problème de Kantorovich. Le but de cette régularisation est de pénaliser l'annulation de γ sur $Supp(\alpha) \times Supp(\beta)$. Cette régularisation résulte en un algorithme itératif très simple pour calculer la solution γ^*. C'est l'algorithme de Sinkhorn.

L'étude du transport optimal déséquilibré n'est pas nouveau. Cependant, dans la littérature existante, il avait été étudié dans des cas bien spécifiques qu'il est difficile de généraliser. Cela est l'une des motivations de l'étude que nous allons présenter dans les pages suivantes. L'approche plus générale que nous présentons permet d'avoir une vision globale tout en gardant un algorithme très efficace.

2.3 Contribution de l'article

La principale contribution de l'article est son algorithme pour résoudre le problème du transport optimal déséquilibré. Cet algorithme propose une solution globale à toutes les problématiques de transport optimal. Cela est possible grâce à une reformulation de l'objectif globale du transport optimal qui unifie le problème du flow de gradients, du barycentre de Wasserstein et du transport déséquilibré. Une preuve de la validité de l'algorithme sera présentée ainsi que des techniques pour assurer sa stabilité.

2.4 Notations

Si X est un espace topologique, et p un nombre réel, $\mathcal{M}_+^p(X)$ est l'espace des mesures de radon positives sur X de masse p. $\mathcal{M}_+(X)$ désigne tout simplement l'espace des mesures de radon positives sur X. Si T est une application de l'espace mesurable $(X, \mathcal{A})$ vers l'espace mesurable $(Y, \mathcal{B})$ et μ est une mesure sur X, alors $T_\sharp\mu$ est la mesure sur Y définie par

$$\forall B \in \mathcal{B}, T_\sharp\mu(B) = \mu(T^{-1}(B)). \tag{1}$$

On note la divergence de Kullback-Leibler entre deux mesures μ et ν sur X par

$$\mathrm{KL}(\mu|\nu) = \int_X \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right)\mathrm{d}\mu(x) + \int_X \mathrm{d}\nu(x) - \mathrm{d}\mu(x) \tag{2}$$

avec la convention $0\log(0/0) = 0$. Si μ n'admet pas de densité $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ par rapport à ν, $\mathrm{KL}(\mu|\nu) = \infty$.

Si r, s de $(X \times Y, \mathrm{d}x\,\mathrm{d}y)$ vers $\mathbb{R}^n$ sont deux familles de fonctions, alors $P_\sharp^X r$ et $P_\sharp^Y r$ sont les fonctions définies par

$$(P_\sharp^X r)_i(x) = \int_Y r_i(x,y)\,\mathrm{d}y \text{ and } (P_\sharp^Y r)_i(y) = \int_X r_i(x,y)\,\mathrm{d}x \tag{3}$$

De même, leur divergence de Kullback-Leibler est définie par

$$\mathrm{KL}(r|s) = \sum_i \int_{X\times Y} \left[\log\left(\frac{r_i(x,y)}{s_i(x,y)}\right) r_i(x,y) - r_i(x,y) + s_i(x,y)\right]\mathrm{d}x\,\mathrm{d}y \tag{4}$$

Enfin, on note

$$< r|s >= \sum_{i=1}^{n} \int_{X\times Y} r_i(x,y)s_i(x,y)\,\mathrm{d}x\,\mathrm{d}y. \tag{5}$$

Les opérateurs $\odot$ et $\oslash$ sont les opérateurs de multiplications et de division coefficient par coefficient pour les matrices. On rappelle que l'indicatrice d'un ensemble convexe C est la fonction

$$\iota_C(x) = \begin{Bmatrix} 0 & \text{si } x \in C \\ \infty & \text{sinon} \end{Bmatrix}$$

Pour finir, nous présentons un opérateur prééminent dans cet article. Si l'on dispose d'une fonction de divergence D sur un ensemble E et si F est une fonction de E vers $\mathbb{R} \cup \{\infty\}$, alors la l'opérateur de proximité est définit par

$$prox_F^D(x) = \text{argmin}_{y\in E}\{F(y) + D(y|x)\} \tag{6}$$

Fondamentalement, c'est un opérateur qui optimise F en privilégiant le voisinage de l'entrée x.

3 Présentation des résultats

3.1 L'équation du Transport Optimal

Dans cette section nous allons nous activer à donner un formalisme globale à toute la problématique du transport optimal. Nous rappelons brièvement la définition des fonctions de divergence.

Définition 1. Fonction d'entropie

Une fonction $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ est dite fonction d'entropie si elle est semi-continue inférieurement, convexe et si $Supp\phi \subset [0, \infty[$ tout en vérifiant $Supp\phi \cap]0, \infty[\neq \emptyset$.

La vitesse de croissance de ϕ à l'infinie est définie par

$$\phi'_\infty = \lim_{x \mapsto \infty} \frac{\phi(x)}{x} \in \mathbb{R} \cup \{\infty\}$$

Définition 2. ϕ-Divergence

Soit ϕ une fonction d'entropie et pour $\alpha, \beta \in \mathcal{M}(X)$, notons $\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\beta + \alpha^\perp$ la décomposition de Lebesgue de α par rapport à β. La ϕ-divergence $\mathcal{D}_\phi$ est définie par :

$$\mathcal{D}_\phi(\alpha|\beta) = \int_X \phi(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta})\,\mathrm{d}\beta + \phi'_\infty \alpha^\perp(X)$$

Les ϕ-divergences sont un bon outil pour comparer des distributions, en particulier on peut remarquer que dans le cas où $\phi'_\infty = \infty$, $\mathcal{D}_\phi(\alpha|\beta) = \infty$ dès que α n'a pas de densité par rapport à β. $\mathcal{D}_\phi$ est elle -même convexe et semi-faiblement continue inférieurement.

L'équation du transport optimal équilibré peut alors se réécrire pour $\alpha, \beta \in \mathcal{M}_+(X) \times \mathcal{M}_+(Y)$

$$\inf_{\gamma \in \mathcal{M}_+(X \times Y)} \int_{X \times Y} c\,\mathrm{d}\gamma + \iota_=(P^X_\sharp \gamma|\alpha) + \iota_=(P^Y_\sharp \gamma|\beta) \tag{7}$$

avec $\iota_=(\mu|\nu) = \mathcal{D}_{\iota_{\{1\}}}(\mu|\nu) = \begin{cases} 0 & \text{si } \mu = \nu \\ \infty & \text{sinon} \end{cases}$. Les deuxième et troisième termes de l'expression servent à pénaliser toute divergence entre $P^X_\sharp \gamma$ et α et entre $P^Y_\sharp \gamma$ et β.

L'avantage de cette nouvelle formulation des contraintes est qu'elle n'implique pas intrinsèquement que $\alpha(X) = \beta(Y)$. On peut alors l'utiliser pour exprimer plus généralement le problème dans le cadre déséquilibré en

$$\inf_{\gamma \in \mathcal{M}_+(X \times Y)} \int_{X \times Y} c\,\mathrm{d}\gamma + \mathcal{D}_{\phi_1}(P^X_\sharp \gamma|\alpha) + \mathcal{D}_{\phi_2}(P^Y_\sharp \gamma|\beta) \tag{8}$$

De manière analogue au cas équilibré, cette quantité peut permettre de définir un analogue à la distance de Wasserstein.

Définition 3. Distance de Wasserstein-Fisher-Rao

Lorsque $X = Y$, que $\phi_1 = \phi_2 = \lambda\phi_{KL}$ avec $\lambda \geq 0$, et que $c(x,y) = -\log\left(\cos^2(d(x,y) \wedge \frac{\pi}{2}\right)$ $WFR_\lambda(\alpha|\mu)$ est défini par la racine carrée du minimum (8).

$$\phi_{KL}(s) = \begin{cases} s\log(s) - s + 1 & \text{si } s > 0 \\ 1 & \text{si } s = 0 \\ \infty & \text{sinon} \end{cases}$$

Proposition 4.

WFR_λ définie bien une distance pour $0 \leq \lambda \leq 1$ (dégénérée si $\lambda = 0$).

Nous ne sommes plus qu'à un pas de pouvoir aussi réécrire le problème des barycentres. Soient $(\alpha_k)_{k=1}^n \in \mathcal{M}_+(X)^n$ associées à des coûts $(c_k)_{k=1}^n$ définies sur $X \times Y$. On veut trouver une distribution "moyenne" $\sigma \in \mathcal{M}_+(Y)$, c'est à dire

$$\inf_{\sigma\in\mathcal{M}_+(Y)} \inf_{\substack{(\gamma)_{k=1}^n \in \\ \mathcal{M}_+(X\times Y)^n}} \sum_k \int_{X\times Y} c_k \,\mathrm{d}\gamma_k + \mathcal{D}_{\phi_1,k}(P^X_\sharp\gamma_k|\alpha_k) + \mathcal{D}_{\phi_2,k}(P^Y_\sharp\gamma_k|\sigma)$$

Si l'on regarde de manière peu formelle, cette expression peut se réécrire grâce à ce qui précède

$$\inf_{\sigma\in\mathcal{M}_+(Y)} \sum_k \text{"distance"}\ (\alpha_k, \sigma)$$

ce qui traduit bien ce qu'on attend de l'équation. On peut bien entendu envisager d'ajouter des poids $(\mu_k)_{k=1}^n$ si on le voulait. Maintenant, si on inverse les deux infimums, on obtient

$$\inf_{\substack{(\gamma)_{k=1}^n \in \\ \mathcal{M}_+(X\times Y)^n}} \left\{ \sum_k \int_{X\times Y} c_k \,\mathrm{d}\gamma_k + \mathcal{D}_{\phi_1,k}(P^X_\sharp\gamma_k|\alpha_k) + \inf_{\sigma\in\mathcal{M}_+(Y)} \sum_k \mathcal{D}_{\phi_2,k}(P^Y_\sharp\gamma_k|\sigma) \right\} \tag{9}$$

Dans les deux cas vus précédemment, une formule générale se déssine. C'est la suivante :

$$\inf_{\gamma\in\mathcal{M}_+(X\times Y)} \mathcal{J}(\gamma) \text{ avec } \mathcal{J}(\gamma) =< c, \gamma > + \mathcal{F}_1(P^X_\sharp\gamma) + \mathcal{F}_2(P^Y_\sharp\gamma) \tag{10}$$

$\mathcal{F}_1$ et $\mathcal{F}_2$ sont deux fonctionnelles semi-continues inférieurement. Cette formule marcherait aussi dans le cadre du flow de gradient, mais cela ne sera pas traité dans cet exposé.

3.2 Regularisation entropique

On munit X et Y de mesures de références $\mathrm{d}x$ et $\mathrm{d}y$. On introduit de nouveau de l'entropie $\mathcal{H}$ qui va jouer cette fois un rôle régularisateur.

$$\mathcal{H}(\gamma) = \sum_k \mathcal{D}_{\phi_{KL}}(\gamma_k | \,\mathrm{d}x\,\mathrm{d}y) = \sum_k \int_{X\times Y} r_k(\log(r_k) - 1)\,\mathrm{d}x\,\mathrm{d}y + const$$

où les r_k sont les densités de γ_k par rapport à $\mathrm{d}x\,\mathrm{d}y$. Cela permet de voir l'importance du choix de $\mathrm{d}x\,\mathrm{d}y$ car il faut que les γ_k soient absolument continues par rapport à elle. En particulier, il faut que $Supp\,\mathrm{d}x\,\mathrm{d}y = X \times Y$. Dans la suite nous négligerons la constante qui n'apporte rien à l'équation. Enfin, en introduisant un paramètre de régularisation ϵ. Nous allons minimiser

$$\inf_{\gamma\in\mathcal{M}_+(X\times Y)^n} \mathcal{J}(\gamma) + \epsilon\mathcal{H}(\gamma) \tag{11}$$

ce qui peut se réécrire

$$\inf_{\gamma\in\mathcal{M}_+(X\times Y)^n} \mathcal{F}_1(P^X_\sharp\gamma) + \mathcal{F}_2(P^Y_\sharp\gamma) + \epsilon\sum_k \mathrm{KL}(\gamma_k | e^{-c_k/\epsilon}\,\mathrm{d}x\,\mathrm{d}y) \tag{12}$$

Remarquons que résoudre (11) revient à calculer $prox^{\mathrm{KL}}_{\mathcal{J}/\epsilon}(\mathrm{d}x\,\mathrm{d}y)$.

Le rôle de régularisation de ϵ peut être compris comme suit. Dans la quantité à minimiser (11), il y a deux quantités qui s'opposent. ϵ va d'une certaine manière décider de laquelle de ces deux influences l'emportera sur le tout. Ainsi lorsque $[\epsilon \longrightarrow 0]$ l'influence de l'entropie disparait et la suite de solutions γ_ϵ de (11) converge vers la solution de (1) ayant une entropie maximale. D'autre part, lorsque $[\epsilon \longrightarrow \infty]$, l'influence de l'entropie l'emporte et la séquence de solutions γ_ϵ de (11)converge vers $\mathrm{d}x\,\mathrm{d}y$.

Proposition 5. Convergence

Considerons le cas où $n = 1$ et où X et Y sont des ensembles finis. Supposons que (10) admette au moins une solution absolument continue par rapport à $\mathrm{d}x\,\mathrm{d}y$. Soit $(\epsilon_k)_k$ une suite de réels strictement positifs qui converge vers 0. Alors la suite des solutions γ_k de (11) converge vers la solution de (10) avec une entropie maximale ($\mathcal{H}(\gamma)$ minimale).

L'hypothèse sur l'absolue continuité permet d'éviter que $\mathcal{D}_{\phi_{KL}}(\gamma | \,\mathrm{d}x\,\mathrm{d}y) \neq \infty$, ce qui permet de bien réaliser la minimisation.

Débarrassons nous donc des mesures et raisonnons en terme de densités en supposant que $\mathrm{d}x\,\mathrm{d}y$ le permette. Alors (12) peut se réécrire :

$$\min_{r\in L^1(X\times Y)^n} F_1(P^X_\sharp r) + F_2(P^Y_\sharp r) + \epsilon\sum_k \mathrm{KL}(r_k | e^{-c_k/\epsilon}) \tag{13}$$

avec $F_1(s) = \mathcal{F}_1(s\,\mathrm{d}x)$, $F_2(s) = \mathcal{F}_2(s\,\mathrm{d}y)$ et $K = (e^{-c_k/\epsilon})_{k=1}^n$. On a remplacé inf par min car on a supposé que la solution existe. Nous résumons plus bas l'ensemble des hypothèses faites.

Définition 6. Hypothèses

Nous faisons les hypothèses suivantes :
- (10) admet au moins une solution absolument continue par rapport à $\mathrm{d}x\,\mathrm{d}y$
- $\mathrm{d}x$ et $\mathrm{d}y$ sont des mesures de probabilité.
- F_1 et F_2 sont faiblement semi continues inférieurement et convexes sur $L^1(X)^n$ et $L^1(Y)^n$ respectivement
- $K \in L^\infty_+(X \times Y)^n$

Maintenant, un des principaux résultats de l'article.

Théorème 7. Dualité

Le dual du problème (13) est

$$\sup_{\substack{u\in L^\infty(X)^n \\ v\in L^\infty(Y)^n}} -F_1^*(-u) - F_2^*(-v) - \epsilon < e^{(u\oplus v)/\epsilon} - 1, K > \tag{14}$$

La solution de cette (14) et de (13) sont égales (dualité forte) et le minimum (13) est atteint en un unique r. De plus, u et v maximisent (14) si et seulement si

$$\left\{ \begin{array}{l} -u \in \partial F_1(P_\sharp^X r) \\ -v \in \partial F_2(P_\sharp^Y r) \end{array} \right\} \text{ et } r_k(x,y) = e^{u_k(x)/\epsilon} K_k(x,y) e^{v_k(y)/\epsilon} \tag{15}$$

C'est un théorème très fort qui nécessite la convexité de F_1 et de F_2. Il nous donne un moyen d'accéder à la densité recherchée à partir de la résolution du problème dual. Notons que cette technique de calcul est très récurrente dans tous les problèmes d'optimisation et donne lieu à des résolutions closes. L'article a réussi à mettre au point un algorithme itératif similaire à Sinkhorn et très simple.

3.3 Algorithme

La reformulation du problème obtenue en (14) nous permet de nous ramener à un cadre d'utilisation de l'algorithme de Dykstra. Les solutions recherchées sont alors obtenues par maximisation successives :

$$\left\{ \begin{array}{l} u^{(l+1)} = \mathrm{argmax}_{u\in L^\infty(X)^n} - F_1^*(-u) - \epsilon < e^{u/\epsilon}, \mathcal{K} e^{v^{(l)}/\epsilon} >_X \\ v^{(l+1)} = \mathrm{argmax}_{v\in L^\infty(Y)^n} - F_2^*(-v) - \epsilon < e^{v/\epsilon}, \mathcal{K}^T e^{u^{(l+1)}/\epsilon} >_Y \end{array} \right\} \tag{16}$$

où pour $a : X \longrightarrow [0,\infty[^n$ et $b : Y \longrightarrow [0,\infty[^n$ mesurables,

$$(\mathcal{K}b)_k(x) = \int_Y K_k(x,y) b_k(y)\,\mathrm{d}y \text{ et } (\mathcal{K}^T a)_k(y) = \int_X K_k(x,y) a_k(x)\,\mathrm{d}x \tag{17}$$

et en remarquant grâce à Fubini-Tonelli que :

$$< e^{(u\oplus v)/\epsilon}, K >_{X\times Y} = < e^{u/\epsilon}, \mathcal{K} e^{v/\epsilon} >_X = < e^{v/\epsilon}, \mathcal{K}^T e^{u/\epsilon} >_Y . \tag{18}$$

Le Théorème suivant va fixer les conditions de validité de cet algorithme ainsi que la manière dont on peut l'implémenter concrètement. Mais avant, il nous faut encore définir une notion

Définition 8. Intégrandes normales et fonctionnelles à noyaux

Une fonction $f : X \times \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{\infty\}$ est appelée intégrande normale si $x \in X \longmapsto \text{epigraphe} f(x, \dot{})$ est à valeurs dans un fermé et mesurable. Une fonctionnelle à noyau convexe est une application $F : L^1(X)^n \longrightarrow \mathbb{R} \cup \{\infty\}$ de la forme $F(s) = \int_X f(x, s(x))\,\mathrm{d}x$ où f est une intégrande normale et $f(x, \dot{})$ est convexe pour tout $x \in X$. Si de plus pour tout $x \in X$, $f(x, \dot{})$ est positive, possède un support inclus dans $[0, \infty[^n$ et s'il existe $s \in L^1(X)^n$ tel que $F(s) < \infty$, alors F est dite admissible.

Théorème 9.

Supposons que F_1 et F_2 soient des fonctionnelles à noyaux admissibles associées aux intégrandes normales f_1 et f_2. Supposons que pour tout $x, y \in X \times Y$, il existe s_1, s_2 à coordonnées strictement positives telles que $f_1(x, s_1) < \infty$ et $f_2(y, s_2) < \infty$ et que K est positive. Si nous définissons $a^{(0)} = 1, v^{(0)} = 0$ et la suite récurrente $(a^{(l)}, b^{(l)})$ par

$$\left\{ \begin{array}{l} a^{(l+1)} = \dfrac{\mathrm{prox}^{\mathrm{KL}}_{F_1/\epsilon}(\mathcal{K} b^{(l)})}{\mathcal{K} b^{(l)}} \\ b^{(l+1)} = \dfrac{\mathrm{prox}^{\mathrm{KL}}_{F_2/\epsilon}(\mathcal{K}^T a^{(l+1)})}{\mathcal{K}^T a^{(l+1)}} \end{array} \right\} \tag{19}$$

alors :

- La suite des maximaux alternés $(u^{(l)}, v^{(l)})$ définie en (16) existe de manière unique.
- Pour tout $l \in \mathbb{N}$, $(a^{(l)}, b^{(l)}) = (e^{u^{(l)}/\epsilon}, e^{v^{(l)}/\epsilon})$.

Ce théorème est un peu le bijou de l'article qui utilise toute la théorie introduite jusqu'ici. Son intérêt pratique est aussi à souligner car il permet de calculer la suite u et v simplement. Enfin un dernier théorème de convergence de l'algorithme.

Théorème 10. Convergence

Sous les mêmes hypothèses que (3.3), si l'équation (19) admet un point fixe (a, b) tel que $(\log a, \log b) \in L^\infty(X)^n \times L^\infty(Y)^n$, alors $(\epsilon \log a, \epsilon \log b)$ est l'unique solution de (14) et la fonction r définie par $r_k(x, y) = a_k(x) K_k(x, y) b_k(y)$ est l'unique solution de (13).

Ce théorème assez intuitif, vient résumer l'ensemble des efforts effectués jusque là et donner lieu à l'algorithme de la figure 1. L'algorithme qu'on y lit contient quelques techniques ajoutées artificiellement afin de garantir une stabilité dans

Algorithm 2 Scaling algorithm with stabilization

function ScalingAlgo2($\text{proxdiv}_{F_1}, \text{proxdiv}_{F_2}, \mathbf{C}, \mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}, \varepsilon$)
  $(\tilde{\mathbf{b}}, \mathbf{u}, \mathbf{v}) \leftarrow (\mathbb{1}_J, 0_I, 0_J)$
  $\tilde{\mathbf{K}}_{ij} \leftarrow \exp(-\mathbf{C}_{ij}/\varepsilon)$ ▷ for all i, j.
  repeat
    $\tilde{\mathbf{a}} \leftarrow \text{proxdiv}_{F_1}(\tilde{\mathbf{K}}(\tilde{\mathbf{b}} \odot \mathrm{d}\mathbf{y}), \mathbf{u}, \varepsilon)$
    $\tilde{\mathbf{b}} \leftarrow \text{proxdiv}_{F_2}(\tilde{\mathbf{K}}^{\mathrm{T}}(\tilde{\mathbf{a}} \odot \mathrm{d}\mathbf{x}), \mathbf{v}, \varepsilon)$
    if a component of $|\log \tilde{\mathbf{a}}|$ or $|\log \tilde{\mathbf{b}}|$ is "too big" **then**
      $(\mathbf{u}, \mathbf{v}) \leftarrow (\mathbf{u} + \varepsilon \log \tilde{\mathbf{a}}, \mathbf{v} + \varepsilon \log \tilde{\mathbf{b}})$
      $\tilde{\mathbf{K}}_{ij} \leftarrow \exp((\mathbf{u}_i + \mathbf{v}_j - \mathbf{C}_{i,j})/\varepsilon)$ ▷ for all i, j.
      $\tilde{\mathbf{b}} \leftarrow \mathbb{1}_J$
    end if
  until stopping criterion
  return $(\tilde{\mathbf{a}}_i \tilde{\mathbf{K}}_{i,j} \tilde{\mathbf{b}}_j)_{i,j}$ ▷ The primal optimizer
end function

Figure 1 – Algorithme pour des mesures discrètes. Image tirée de l'article

les calculs.

3.4 Expérience

Pour finir, j'ai voulu présenter une petite illustration de ce que réalise l'algorithme. On se place dans le cas $2-d$ où les points de X sont 300 points pris indépendamment dans le carré de côté 1 et centré en l'origine. Les 200 points de Y sont prix uniformément dans la couronne comprise entre les cercles de rayon 0.8 et 0.2 centrés en l'origine. Les distributions $\mathrm{d}x$ et $\mathrm{d}y$ sont des sommes de diracs en tous les points avec une même masse. Pour la divergence, j'ai pris la $\mathcal{D}_{\lambda\phi_{\mathrm{KL}}}$ avec $\lambda = 3$. Nous avons afficher l'erreur en échelle logarithmique au cours des itérations en figure 2 et le couplage obtenu en figure 3. J'ai implémenté l'algorithme moi-même.

4 Conclusion et LIEN avec le cours

L'une des prouesses remarquables est la formule de réunification des problématiques du transport optimale. Elle donne une vision d'ensemble très pertinente. La résolution de l'équation est rigoureuse et l'algorithme qui en découle est simple et performant. L'algorithme tel que je l'ai implémenté a une complexité constante car le nombre d'itération est fixé d'avance comme suggéré dans l'article. Seulement, il aurait manqué une petite justification de la définition (3.3) en montrant que les fonctionnelles F_1 et F_2 que nous utilisons sont bien admissibles.

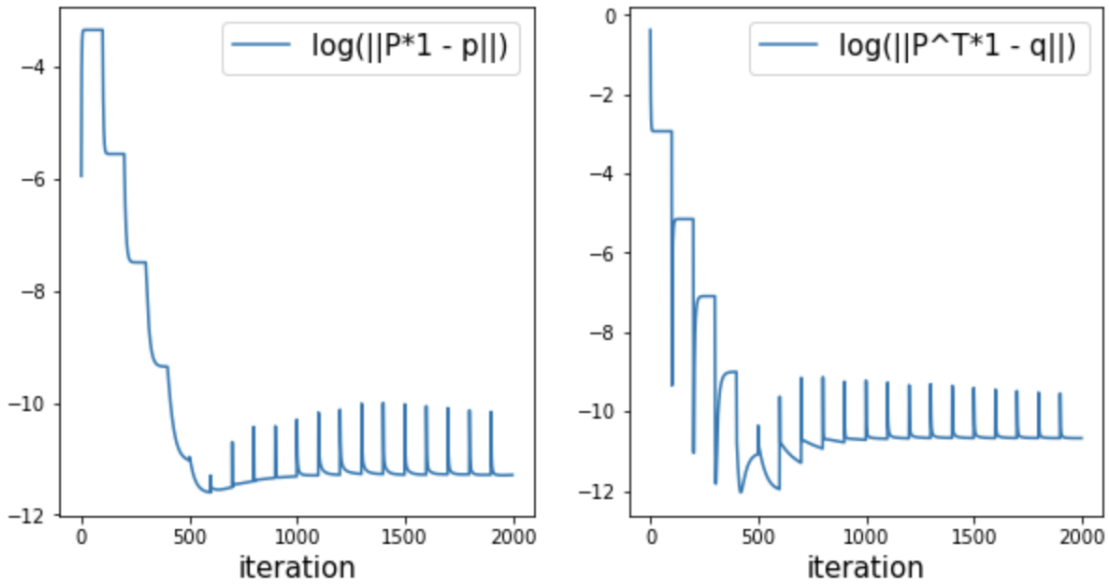

FIGURE 2 – Erreur affichée en echelle logarithmique

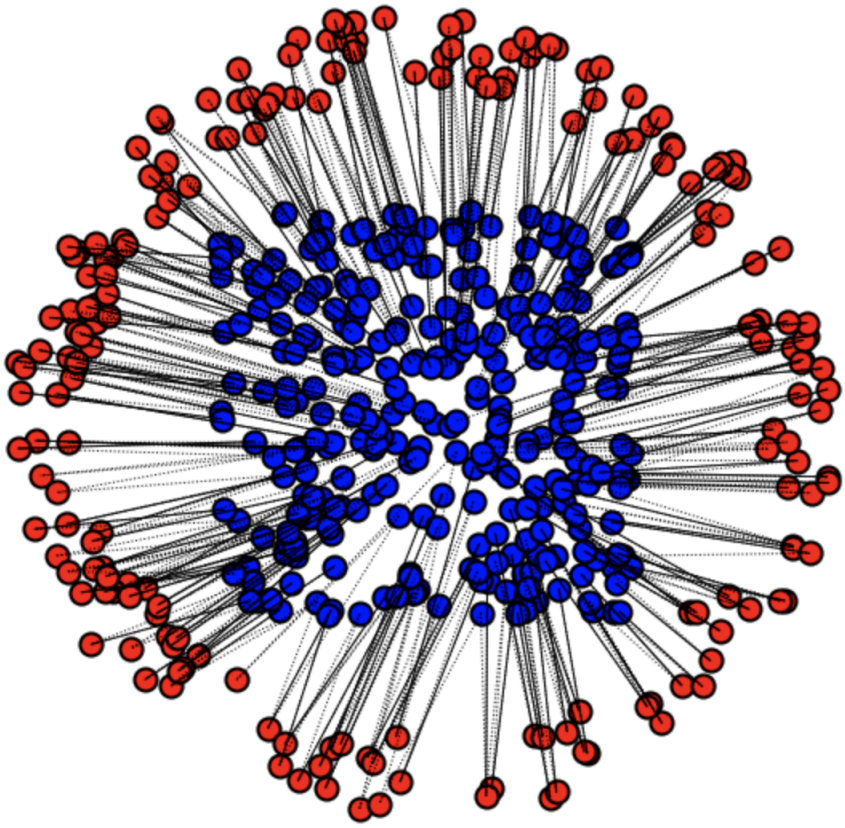

FIGURE 3 – Illustration du couplage obtenu

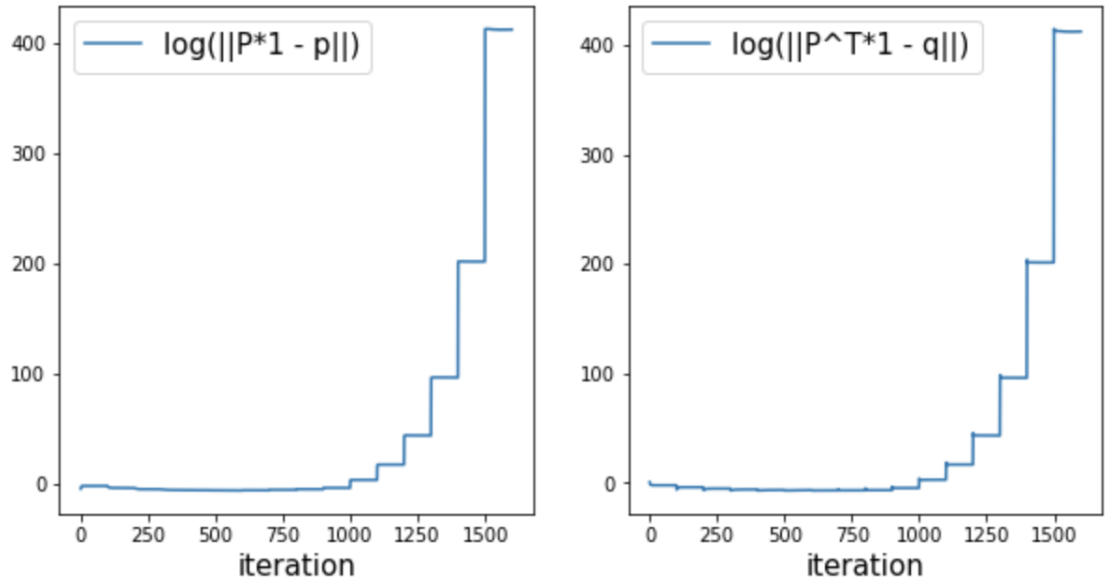

FIGURE 4 – Erreur obtenue lorsqu'on donne des poids nuls à la moitié des éléments de X et de Y. Confirme les remarques faites sur le choix de dx et dy

L'ensemble des équations étudiées dans cet article avaient déjà été résolues dans le cours dans le cadre du transport optimal équilibré. L'algorithme de Sinkhorn avait été obtenu en réalisant une résolution par régularisation entropique puis par dualité. L'algorithme obtenu dans cet article ressemble en tous points à Sinkhorn car il est lui aussi obtenu par régularisation entropique et dualité. Il est à noter que l'algorithme ici présenté est tout à fait l'algorithme de Sinkhorn lorsque X et Y ont même masse. Cet article généralise donc Sinkhorn. La distance WFR constitue un parfait analogue de la distance de Wasserstein.

5 Bibliographie

La bibliographie est essentiellement constituée de :

— Gabriel PEYRE (https ://optimaltransport.github.io/slides-peyre/CourseOT.pdf)
— Scaling Algorithm for unbalanced optimal transport problems. Lénaïc CHIZAT, Gabriel PEYRE, Bernhard SCHMITZER, et François-Xavier VIALARD