Le transport optimal numérique

Guillaume Lambert - Leonardo Martins Bianco

25 Janvier 2021

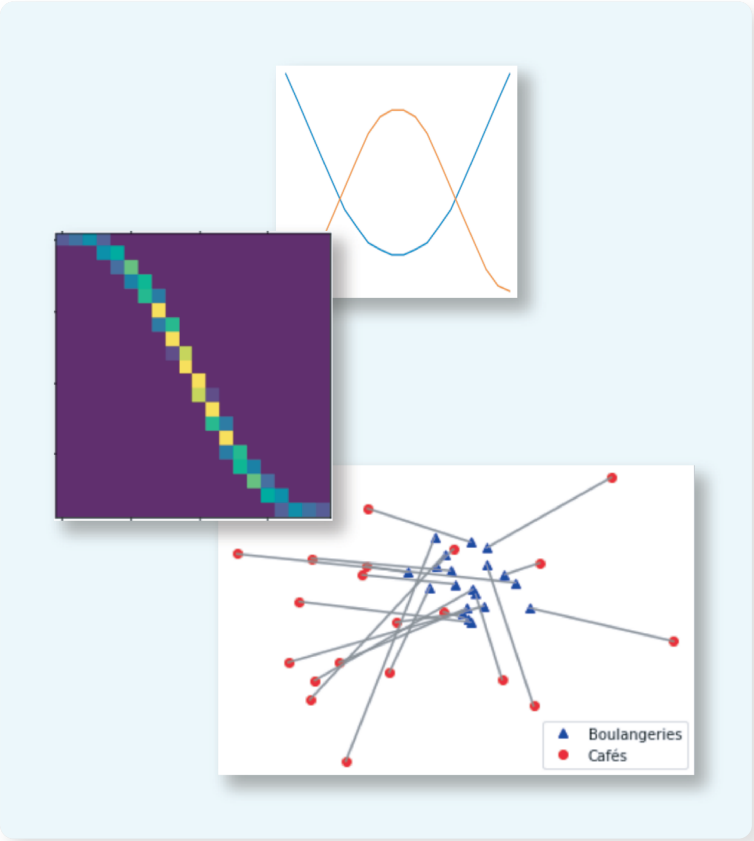


Table des matières

1 Introduction

Dans ce projet, nous étudions le problème de transport optimal et différentes méthodes de résolution numériques de ce problème. Il consiste à chercher le moyen le moins coûteux (en temps, distance ou argent par exemple) pour transporter des objets entre un ensemble de points de départ et un ensemble de points d'arrivée.

Ce concept a plusieurs applications. Dans ce projet, on a appliqué ces techniques au problème de matching entre boulangeries-cafés. Mais on peut aussi par exemple appliquer cette théorie à la physique [1], aux problèmes de proximité sémantique des mots [2] ou aux problèmes du traitement du signal comme la normalisation de niveau de gris [3]. Le transport optimal est aussi important dans l'étude des EDPs d'évolution, qui modélisent la dynamique des populations en biologie et sociologie, et dans la mécanique des fluides [4]. Finalement, il est important de noter que plusieurs méthodes du transport optimal ont été développées pour l'économie.

L'introduction du problème de transport optimal sera faite en première partie. Ensuite, nous résoudrons ce problème, qui est un problème de programmation linéaire, par l'algorithme du simplex. Cette méthode de résolution étant très coûteuse en termes de calculs, elle n'est pas viable pour des données de grandes dimensions. C'est pourquoi on étudiera 2 méthodes de régularisation du problème : une régularisation du problème dual grâce à la fonction $log - sum - exp$ et une régularisation entropique du problème primal. Ces régularisations nous donneront des propriétés intéressantes. Tout d'abord, les problèmes régularisés seront une approximation du problème de transport dont la fonction-objectif est convexe. De plus, ils seront sans contraintes. On pourra donc résoudre ces problèmes régularisés par un algorithme de gradient moins coûteux en calculs et plus simple d'implémentation. On comparera les résultats trouvés avec ces 2 stratégies avec ceux trouvés avec la programmation linéaire (simplex).

Finalement, on va réaliser une descente de gradient directement dans l'espace de Wasserstein, via une formulation variationnelle (c'est-à-dire, via la minimisation d'une fonctionnelle d'énergie). Pour conclure, on reprendra cette étude avec la divergence de Sinkhorn, une quantité plus adaptée à ce problème.

2 Mise en forme du problème

Le problème du transport optimal a été introduit par le mathématicien français Gaspard Monge dans [5]. L'objectif était de "transporter des terres d'un lieu dans un autre". À l'époque, ce type de problème était important pour des applications militaires, comme la construction des fortifications. Notons que Monge a introduit le problème, mais il ne l'a pas résolu. Pour une discussion sur ce mémoire, on fait référence à [6].

Après presque deux siècles, le mathématicien soviétique Leonid Kantorovich a fait des progrès pour résoudre ce problème. C'est pourquoi on appelle le problème central de la théorie le problème de Monge-Kantorovich. On l'introduit maintenant, en faisant l'analogie au transport de pains au chocolat des boulangeries vers des cafés.

Imaginons que l'on a N boulangeries qui produisent des pains au chocolat, et N cafés qui demandent ces produits. La production des pains au chocolat est décrite par $\mu \in \mathbb{R}^+$ et la consommation des cafés par $\nu \in \mathbb{R}^+$. On suppose que la quantité totale de la production et de la consommation sont égales :

$$\sum_{i=1}^{N} \mu_i = \sum_{j=1}^{N} \nu_j = 1$$

Cependant, il y a des coûts pour faire ce transport. On peut imaginer par exemple que le coût est proportionnel à la distance entre la boulangerie et le café. On note par C_{ij} le coût de transport entre la i-ème boulangerie et le j-ème café.

Notre but est donc de trouver un couplage optimal $\gamma \in \mathbb{R}_+^{N \times N}$, qui indique comment la production de la i-ème boulangerie est répartie entre tous les cafés. On dira qu'un couplage est déterministe si pour toute boulangerie, l'intégralité des pains au chocolat vont à un seul café.

Mathématiquement, le problème de Monge-Kantorovich s'écrit sous la forme

$$\min \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} C_{ij}\gamma_{ij} | \gamma \in \Pi(\mu, \nu) \right\} \quad (\mathcal{MK})$$

où

$$\Pi(\mu, \nu) := \left\{ \gamma \in \mathbb{R}_+^{N \times N} | \sum_{j=1}^{N} \gamma_{ij} = \mu_i \forall i \in I \text{ et } \sum_{i=1}^{N} \gamma_{ij} = \nu_j \forall j \in J \right\}$$

Passons maintenant aux méthodes de résolution de ce problème.

3 Etude du problème par la programmation linéaire

Dans cette première partie, on souhaite étudier le probème de Monge-Kantorovich $(\mathcal{MK})$ par la programmation linéaire. Pour ce faire, on va adopter la stratégie suivante :

— Mise sous forme standard-primale

— Mise sous forme duale

— Applications :

  ⋆ Résolution d'un exemple en 1D

  ⋆ Résolution par une permutation optimale, d'un exemple en 1D puis un second en 2D

On utilise l'algorithme du simplex dans les résolutions des différents exemples.

3.1 Mise sous forme standard-primale

On souhaite mettre le problème $(\mathcal{MK})$ sous la forme standard-primale

$$\min \{\langle \mathbf{c}, \mathbf{x} \rangle | A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0\} \tag{$\mathcal{P}$}$$

On introduit donc les nouvelles notations suivantes :

— $\mathbf{c} := (C_{11}, C_{12}, ..., C_{1N}, C_{21}, ..., C_{N1}, ..., C_{NN})$ un vecteur de taille N^2

— $\mathbf{x} := (\gamma_{11}, \gamma_{12}, ..., \gamma_{1N}, \gamma_{21}, ..., \gamma_{N1}, ..., \gamma_{NN})$ un vecteur de taille N^2

— $\mathbf{b} := (\mu_1, ..., \mu_N, \nu_1, .., \nu_N)$ un vecteur de taille $2N$

On commence par mettre la fonction-objectif $\gamma \mapsto \sum_{i=1}^{N} \sum_{j=1}^{N} C_{ij}\gamma_{ij}$ de $(\mathcal{MK})$ sous la forme du produit scalaire $\langle \mathbf{c}, \mathbf{x} \rangle$:

$$\begin{aligned} \langle \mathbf{c}, \mathbf{x} \rangle &= \sum_{l=1}^{N^2} \mathbf{c}_l \mathbf{x}_l \\ &= \sum_{m=0}^{N-1} \sum_{n=1}^{N} \mathbf{c}_{mN+n} \mathbf{x}_{mN+n} \end{aligned}$$

Or $\mathbf{c}_{mN+n} = C_{m+1,n}$ et $\mathbf{x}_{mN+n} = \gamma_{m+1,n}$. Donc

$$\begin{aligned} \langle \mathbf{c}, \mathbf{x} \rangle &= \sum_{m=0}^{N-1} \sum_{n=1}^{N} C_{m+1,n}\gamma_{m+1,n} \\ &= \sum_{m=1}^{N} \sum_{n=1}^{N} C_{m,n}\gamma_{m,n} \quad \text{par le changement de variable } m \mapsto m-1 \end{aligned}$$

On met ensuite les contraintes sous la forme de l'équation matricielle $A\mathbf{x} = \mathbf{b}$ avec $\mathbf{x} \geq 0$ et A une matrice de taille $2N \times N^2$.

Premièrement, on a bien $\mathbf{x} \geq 0$ car $\gamma \in \mathbb{R}_+^{N \times N}$. Ensuite, on construit la matrice A : on la décompose en $A = [A_1; A_2]$ et on construit A_1, la matrice des N premières lignes de A :

$$\text{Soit } i \in \{1, ..., N\}, \text{ alors } A_{ij} = \begin{cases} 1 & \text{si } j = (i-1)N + k, \text{ avec } k \in \{1, ..., N\} \\ 0 & \text{sinon} \end{cases}$$

$$A_1 = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & & & & & & & & & & & & \vdots \\ \vdots & & & & & & & & & & & & \vdots \\ \vdots & & & & & & & & & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix}$$

Ensuite, on construit A_2, la matrice des N dernières lignes de A :

$$\text{Soit } i \in \{N+1, ..., 2N\}, \text{ alors } A_{ij} = \begin{cases} 1 & \text{si } j = i + (k-2)N, \text{ avec } k \in \{1, ..., N\} \\ 0 & \text{sinon} \end{cases}$$

$$A_2 = \begin{pmatrix} 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 1 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 1 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & & & & & & & & & & & & & & & & \vdots \\ \vdots & & & & & & & & & & & & & & & & \vdots \\ \vdots & & & & & & & & & & & & & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 1 & 0 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 1 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

On construit donc la matrice A et on obtient finalement la forme standard-primale de $(\mathcal{MK})$:

$$\min \{\langle \mathbf{c}, \mathbf{x} \rangle | A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0\} \qquad (\mathcal{P})$$

Ce problème étant sous forme standard, on pourra le résoudre avec l'algorithme du simplex.

3.2 Mise sous la forme duale

On souhaite maintenant construire le dual $(\mathcal{MK}_d)$ de $(\mathcal{MK})$. On a vu que ce dernier peut s'écrire sous la forme standard-primale $(\mathcal{P})$ suivante :

$$\min \{\langle \mathbf{c}, \mathbf{x} \rangle | A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0\} \qquad (\mathcal{P})$$

Le problème dual s'écrit donc

$$\max \left\{\langle \mathbf{b}, \mathbf{y} \rangle | A^\top \mathbf{y} \leq \mathbf{c}\right\} \qquad (\mathcal{MK}_d)$$

avec $\mathbf{y}$ un vecteur de taille $2N$ que l'on décompose en deux vecteurs $\mathbf{u}$ et $\mathbf{v}$ de taille N.

On commence par exprimer la fonction-objectif :

$$\begin{aligned}\langle \mathbf{b}, \mathbf{y} \rangle &= \sum_{i=1}^{2N} \mathbf{b}_i \mathbf{y}_i \\ &= \sum_{i=1}^{N} \mathbf{b}_i \mathbf{y}_i + \sum_{i=N+1}^{2N} \mathbf{b}_i \mathbf{y}_i \\ &= \sum_{i=1}^{N} \mu_i u_i + \sum_{i=1}^{N} \nu_i v_i\end{aligned}$$

ensuite, on exprime les contraintes :

Soit $(i,j) \in I \times J$. On a vu précédemment que $\mathbf{c}_{(i-1)N+j} = C_{i,j}$ avec $(i,j) \in I \times J$. Par la construction du dual, on a aussi que

$$\begin{aligned}C_{i,j} = \mathbf{c}_{(i-1)N+j} \geq \left(A^\top \mathbf{y}\right)_{(i-1)N+j} &= \sum_{k=1}^{2N} \left(A^\top\right)_{(i-1)N+j,k} \mathbf{y}_k \\ &= \sum_{k=1}^{N} \left(A^\top\right)_{(i-1)N+j,k} \mathbf{y}_k + \sum_{k=N+1}^{2N} \left(A^\top\right)_{(i-1)N+j,k} \mathbf{y}_k \\ &= \sum_{k=1}^{N} A_{k,(i-1)N+j} \;\; u_k + \sum_{k=N+1}^{2N} A_{k,(i-1)N+j} \;\; v_{k-N} \\ &= \sum_{k=1}^{N} \mathbf{1}_{i=k} \;\; u_k + \sum_{k=N+1}^{2N} \mathbf{1}_{j=k-N} \;\; v_{k-N} \\ &= u_i + v_j\end{aligned}$$

Le problème dual $(\mathcal{MK}_d)$ se réécrit donc

$$\max \left\{ \sum_{i=1}^{N} \mu_i u_i + \sum_{i=1}^{N} \nu_i v_i \mid \mathbf{u}, \mathbf{v} \in \mathbb{R}^N \; u_i + v_j \leq C_{ij} \; \forall (i,j) \in I \times J \right\} \qquad (\mathcal{MK}_d)$$

On pourra aussi résoudre ce problème avec l'algorithme du simplex. En effet, on peut le mettre sous forme standard : on transforme la fonction-objectif avec $\max(z(x)) = -\min(z(x))$ et on ajoute des variables d'écart pour "activer" les contraintes.

3.3 Applications

3.3.1 Résolution d'un exemple en 1D

Dans cet exemple, on utilise les données suivantes :

— Le domaine étudié est l'intervalle $[0,1]$ discrétisé uniformément par $N = 20$ points. Les boulangeries et les cafés sont donc positionnés aux mêmes points.

— La fonction-coût est la distance au carré : $c(x, y) = |x - y|^2$.
— La fonction de production est $\mu(x) = \mathbf{1}_{[0,1]}(x)$.
— La fonction de consommation est $\nu(x) = \exp(-10(y - 0.5)^2)$.

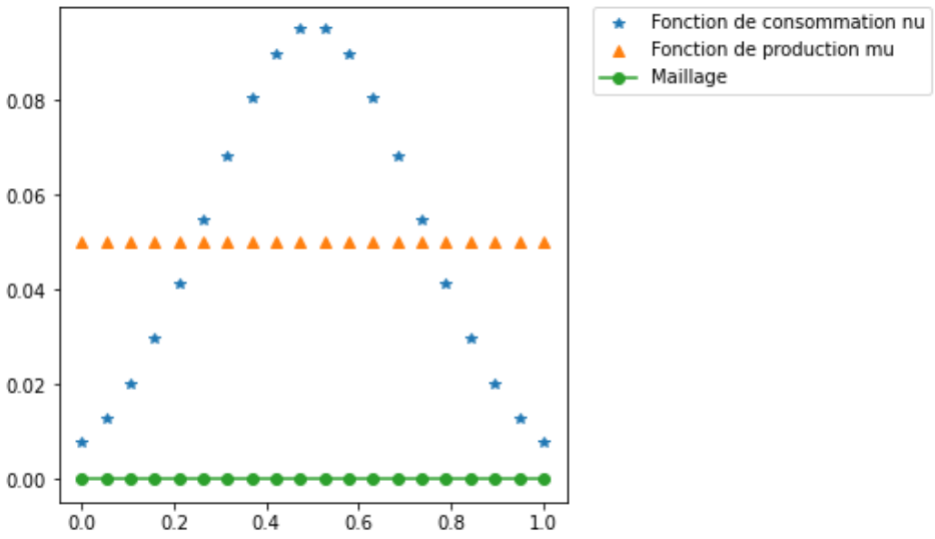


FIGURE 1

On voit que la production est constante dans l'intervalle $[0, 1]$ alors que la consommation est symétrique et centrée en $0,5$. Cela nous donne une approche graphique du problème et une intuition sur les résultats que l'on va calculer.

Les distributions calculées à partir de μ et ν sont normalisées pour que leur masse totale soit 1, ce sont donc des distributions de probabilité.

Résolution de la forme standard-primale :

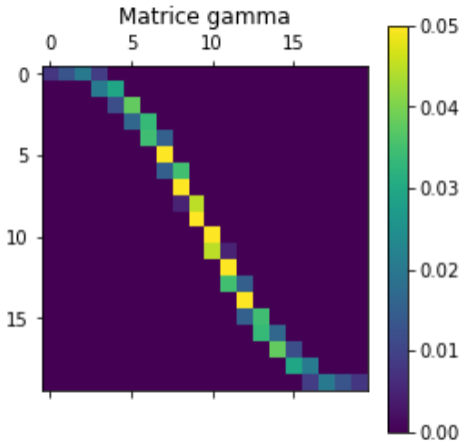


FIGURE 2

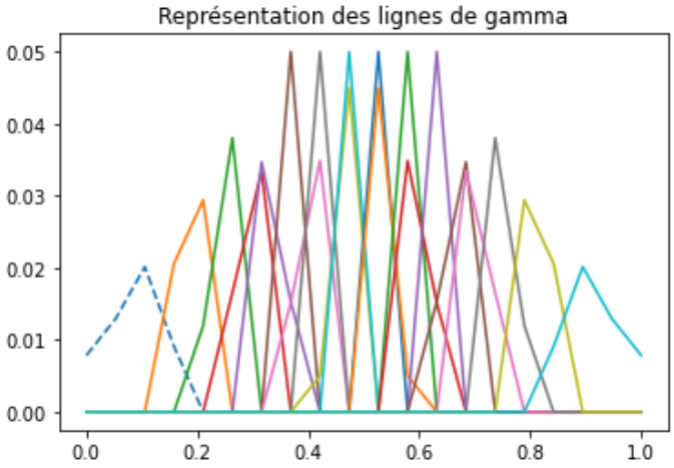


FIGURE 3

Premièrement, on remarque que la symétrie de la production et de la consommation entraîne la symétrie du couplage optimal $\gamma^\star$.

Ensuite, la concentration de la consommation au centre de l'intervalle $[0, 1]$ entraîne des valeurs de $\gamma^\star_{ij}$ importantes au centre de la matrice $\gamma^\star$ (figure 2) et au centre de $[0, 1]$ (figure 3). De plus, on remarque une "attraction" du couplage $\gamma^\star$ vers le centre. Dans la figure 2, on voit une ondulation vers le centre aux extrémités de la ligne des valeurs non nulles du couplage. Dans la figure 3, on voit par exemple que le couplage bleu en pointillé (la première ligne de $\gamma^\star$) va favoriser les cafés à droite (vers le centre). Ce favoritisme vers le centre se voit aussi de manière globale sur chaque ligne du couplage. Enfin, on voit que les 6 boulangeries du centre de l'intervalle $[0, 1]$, où la consommation

est forte, se concentrent sur un unique café chacune.

Concernant la structure de la matrice $\gamma^\star$, on remarque qu'elle n'est pas sparse. On peut aussi interpréter la figure 3 : on voit graphiquement que les productions des boulangeries ne sont pas dédiées chacune à un café. Le couplage n'est donc pas déterministe.

Résolution de la forme duale

Rappelons le problème dual $(\mathcal{MK}_d)$:

$$\max \left\{ \sum_{i=1}^{N} \mu_i u_i + \sum_{i=1}^{N} \nu_i v_i \mid \mathbf{u}, \mathbf{v} \in \mathbb{R}^N \; u_i + v_j \leq C_{ij} \; \forall (i,j) \in I \times J \right\} \qquad (\mathcal{MK}_d)$$

Etant donnée la positivité des distributions μ et ν, on aimerait idéalement avoir $\mathbf{u}$ et $\mathbf{v}$ positifs et les plus grands possibles afin de maximiser la fonction-objectif de $(\mathcal{MK}_d)$. Cependant, la contrainte ne le permet pas. En effet, prenons les couples (i,i), alors $u_i + v_i \leq C_{ii} = |x_i - y_i|^2 = 0$. On peut se dire intuitivement que les v_i et les u_i vont se compenser pour respecter les contraintes.

Pour résoudre ce problème avec l'algorithme du simplex, on utilise le fait que $\max z(x) = -\min(-z(x))$. On va donc chercher à résoudre le problème suivant :

$$\min \left\{ -\sum_{i=1}^{N} \mu_i u_i - \sum_{i=1}^{N} \nu_i v_i \mid \mathbf{u}, \mathbf{v} \in \mathbb{R}^N \; u_i + v_j \leq C_{ij} \; \forall (i,j) \in I \times J \right\}$$

On obtient le résultat suivant

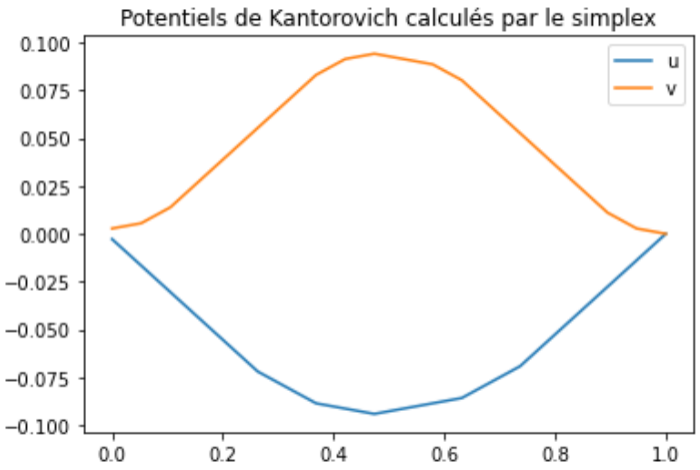


FIGURE 4

Tout d'abord, on observe la "compensation" des u_i par les v_i. Ensuite, on remarque une symétrie et aussi que les valeurs les plus grandes en valeur absolue sont au centre. Ce comportement est causé par la forme de la fonction de consommation. On voit sur la figure 1 qu'elle est symétrique et qu'elle est plus grande que la fonction de production au centre. Ainsi, les contributions $-\nu_i v_i < 0$ sont très

minimisantes pour la fonction-objectif et compensent largement les contributions $-\mu_i u_i > 0$. On peut résumer la stratégie des potentiels de Kantorovich par "miser sur la forte consommation au centre".

Afin de valider nos résultats, on procède à quelques vérifications.

Tout d'abord, on regarde si les potentiels de Kantorovich calculés respectent les contraintes :

$$u_i + v_j \leq C_{ij} \; \forall (i,j) \in I \times J$$

Ensuite, on regarde s'ils vérifient la condition de complémentarité issue de la condition d'optimalité de KKT : $x_i^\star s_i = 0 \quad \forall i \in \{1, ..., N\}$ où $x^\star$ est la solution optimale du problème primal $(\mathcal{P})$ et s la variable d'écart du problème dual $(\mathcal{MK}_d)$. La vérification est faite dans le code.

Enfin, on retrouve le résultat de la dualité forte : si $\mathbf{x}^\star$ est la solution optimale de $(\mathcal{MK})$ et $\mathbf{y}^\star$ est la solution optimale de $(\mathcal{MK}_d)$, alors $\langle \mathbf{c}, \mathbf{x}^\star \rangle = \langle \mathbf{b}, \mathbf{y}^\star \rangle$. Dans notre cas, on obtient :

$$\langle \mathbf{c}, \mathbf{x}^\star \rangle = 0.011112315676793683$$
$$\langle \mathbf{b}, \mathbf{y}^\star \rangle = 0.011112315676793912$$

La différence est égale à 2.310^{-16} et on la considère comme nulle. On a donc bien le résultat de la dualité forte et cela valide nos résultats.

3.3.2 Résolution par une permutation optimale d'un exemple en 1D puis un second en 2D

Dans cet exemple, on utilise les données suivantes :

- Les emplacements des boulangeries et des cafés sont déterminés aléatoirement sur l'intervalle $[0, 1]$ selon la loi uniforme continue sur $[0, 1]$.
- La fonction-coût reste la distance au carré : $c(x, y) = |x - y|^2$.
- Les fonctions production et consommation sont la constante : $\mu(x) = \nu(x) = 1/N$.

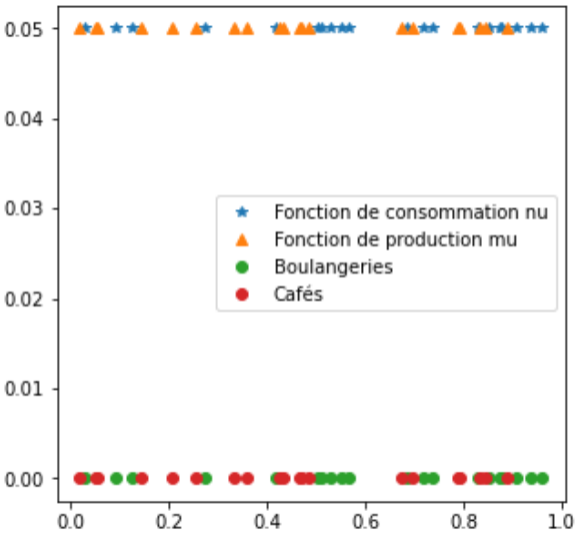


FIGURE 5

Contrairement à l'exemple précédent, la consommation et la production sont égales. Cependant, les boulangeries et les cafés sont placés aléatoirement : chaque résolution du problème est donc unique. On remarque que les distributions μ et ν sont déjà normalisées (ce sont des distributions de la loi uniforme discrète sur l'ensemble discrétisé de $N = 20$ points).

On résout le problème primal et on représente la matrice de couplage optimal γ :

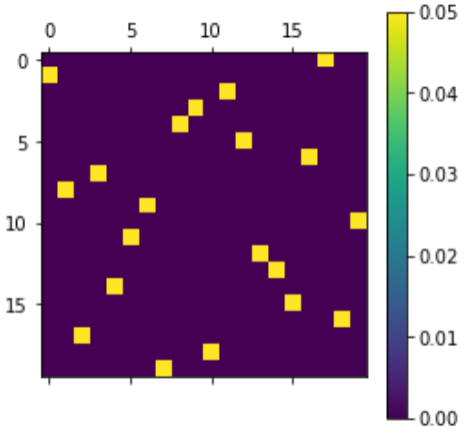


FIGURE 6

On remarque que le couplage optimal $\gamma^\star$ est déterminée par une permutation optimale $\sigma^\star$. En effet, chaque ligne et chaque colonne de $\gamma^\star$ est nulle à l'exception d'un coefficient de valeur $1/N$.

Ainsi, une boulangerie est couplée à un unique café, c'est-à-dire que le couplage est déterministe. Cela peut aussi se voir sur la structure de $\gamma^\star$: c'est une matrice sparse.

On représente ces couples sur le graphique suivant :

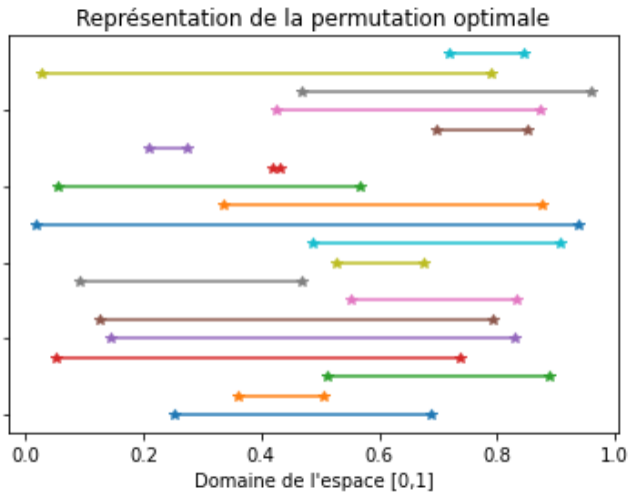


FIGURE 7

Introduisons un nouvel exemple, cette fois ci dans un domaine en 2D. on utilise les données suivantes :

- Les emplacements des boulangeries sont déterminés aléatoirement sur le carré $[0, 1]$: chacune des deux composantes d'un point x_i est déterminée selon la loi uniforme continue sur $[0, 1]$.
- Les emplacements des cafés sont déterminés aléatoirement sur $\mathbb{R}^2$: chacune des deux composantes d'un point y_j est déterminée selon la loi normale centrée réduite.
- La fonction-coût est $c(x_i, y_j) = |x_i^1 - y_j^1|^2 + |x_i^2 - y_j^2|^2$.
- Les fonctions production et consommation sont la constante : $\mu(x) = \nu(x) = 1/N$.

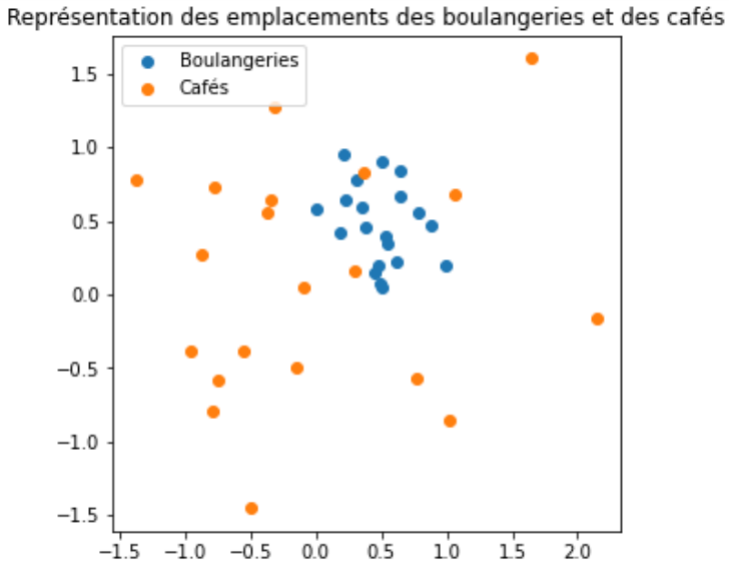


FIGURE 8

On ne représente pas la production et la consommation mais on sait qu'elles sont constantes égales à $1/N = 0,05$. La répartition des boulangeries et des cafés est cohérente avec les lois utilisées pour calculer leur emplacement.

On résout le problème primal et on représente la matrice de couplage optimal $\gamma^\star$:

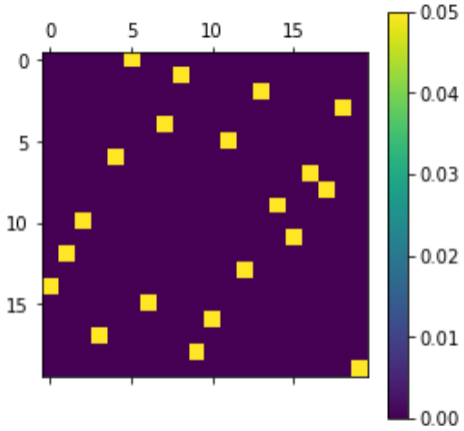


FIGURE 9

Tout comme dans l'exemple précédent, la matrice $\gamma^\star$ est une matrice de permutation. Elle est sparse et le couplage optimal est donc déterministe.

On représente les couples formés sur le graphe suivant :

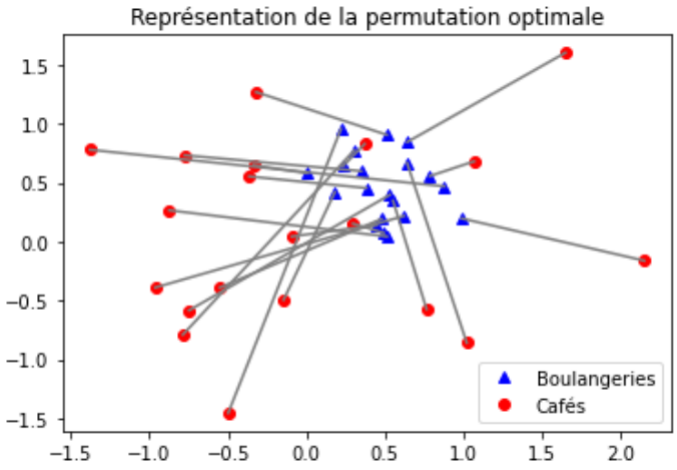


FIGURE 10

Les couplages trouvés dans les figures 7 et 10 semblent visuellement bons et on déduit intuitivement que les 2 solutions trouvées sont optimales.

4 La régularisation par la fonction log-sum-exp

Un problème de l'approche avec l'algorithme du simplex est que cela ne marche bien que pour des problèmes avec beaucoup de points. C'est-à-dire, on a un problème de "scalabilité" quand on utilise le simplex, et la plupart des applications réelles demandent un schéma numérique efficace pour les grandes dimensions. On va appeler un tel schéma une "régularisation" du problème original.

Dans cette deuxième partie, on souhaite régulariser le problème dual $(\mathcal{MK}_d)$ à l'aide de la fonction $log - sum - exp$. On montre ensuite que ce problème régularisé sans contraintes, noté $(\mathcal{MK}_\varepsilon)$, possède les propriétés importantes que l'on recherche : la fonction-objectif de $(\mathcal{MK}_\varepsilon)$ approxime celle de $(\mathcal{MK}_d)$ et elle est convexe. On conclut cette partie par la résolution de l'exemple de la partie 3.3.1 avec le problème $(\mathcal{MK}_\varepsilon)$ par l'algorithme de gradient et la comparaison des résultats obtenus. On adopte donc la démarche suivante :

— Etude de la fonction log-sum-exp

— Mise en forme de $(\mathcal{MK}_d)$ sous une forme régularisée sans contraintes $(\mathcal{MK}_\varepsilon)$

— Mise en application sur l'exemple défini dans la partie 3.3.1

4.1 Etude de la fonction log-sum-exp

Pour mettre $(\mathcal{MK}_d)$ sous une forme sans contrainte et donc utiliser un algorithme de gradient, on va utiliser la fonction log-sum-exp. Elle possède 2 propriétés intéressantes :

— elle est convexe, donc les points stationnaires trouvés par l'algorithme de gradient seront des minima globaux.
— elle possède une propriété de convergence intéressante pour mettre $(\mathcal{MK}_d)$ sous une forme sans contrainte.

Montrons qu'elle est convexe :

On considère la fonction log-sump-exp $f : \mathbb{R}^N \to \mathbb{R}$ définie par

$$f(x) = f((x_1, ..., x_N)) = \log\left(\sum_{i=1}^{N} \exp(x_i)\right)$$

Soient $x, y \in \mathbb{R}^{\mathbb{N}}$ et $t \in [0, 1]$.

$$\begin{aligned}
f(tx + (1-t)y) &= \log\left(\sum_{i=1}^{N} e^{tx_i + (1-t)y_i}\right) \\
&= \log\left(\sum_{i=1}^{N} e^{tx_i} e^{(1-t)y_i}\right) \\
&= \log\left(\sum_{i=1}^{N} a_i^t b_i^{1-t}\right) \text{ avec } a_i = e^{x_i} \text{ et } b_i = e^{y_i} \\
&\leq \log\left(\left(\sum_{i=1}^{N} a_i^{t\cdot\frac{1}{t}}\right)^t \cdot \left(\sum_{i=1}^{N} b_i^{(1-t)\cdot\frac{1}{1-t}}\right)^{1-t}\right) \text{ par l'inégalité de Hölder} \\
&= t\log\sum_{i=1}^{N} a_i + (1-t)\log\sum_{i=1}^{N} b_i \\
&= tf(x) + (1-t)f(y)
\end{aligned}$$

Donc f est bien convexe.

Montrons maintenant la propriété de convergence :

$f_\varepsilon(x) := \varepsilon f(\frac{x}{\varepsilon}) = \varepsilon \log\left(\sum_{i=1}^{N} e^{\frac{x_i}{\varepsilon}}\right)$ converge vers $\max_{i\in\{1,...,N\}}$ quand $\varepsilon \to 0$.

On commence par montrer l'inégalité suivante

$$\max_{i\in\{1,...,N\}} x_i \leq \log\left(\sum_{i=1}^{N} e^{x_i}\right) \leq \max_{i\in\{1,...,N\}} x_i + \log N$$

On a

$$\max_{i\in\{1,...,N\}} e^{x_i} \leq \sum_{i=1}^{N} e^{x_i} \leq N \max_{i\in\{1,...,N\}} e^{x_i}$$

Comme l'exponentielle est croissante, on a $\max_{i\in\{1,\dots,N\}} e^{x_i} = e^{\max_{i\in\{1,\dots,N\}} x_i}$ et donc les inégalités précédentes deviennent

$$e^{\max_{i\in\{1,\dots,N\}} x_i} \leq \sum_{i=1}^{N} e^{x_i} \leq N e^{\max_{i\in\{1,\dots,N\}} x_i}$$

On peut appliquer la fonction logarithmique car $e^{\max_{i\in\{1,\dots,N\}} x_i} > 0$ et grâce à sa croissance, on obtient

$$\max_{i\in\{1,\dots,N\}} x_i \leq \log\left(\sum_{i=1}^{N} e^{x_i}\right) \leq \log(N) + \max_{i\in\{1,\dots,N\}} x_i$$

C'est bien l'inégalité que l'on voulait montrer. On peut maintenant déduire la propriété de convergence recherchée :

On utilise l'inégalité que l'on vient de montrer avec la transformation $x \mapsto x/\varepsilon$:

$$\max_{i\in\{1,\dots,N\}} \frac{x_i}{\varepsilon} \leq f(\frac{x}{\varepsilon}) \leq \log(N) + \max_{i\in\{1,\dots,N\}} \frac{x_i}{\varepsilon}$$

On multiplie ensuite par $\varepsilon > 0$:

$$\max_{i\in\{1,\dots,N\}} x_i \leq f_\varepsilon(x) = \varepsilon f(\frac{x}{\varepsilon}) \leq \varepsilon \log(N) + \max_{i\in\{1,\dots,N\}} x_i$$

Par le théorème des gendarmes, on a bien $f_\varepsilon(x) \to \max_{i\in\{1,\dots,N\}}$ quand $\varepsilon \to 0$.

4.2 Mise en forme de $(\mathcal{MK}_d)$ sous une forme régularisée sans contraintes

Tout d'abord, on peut remarquer que $v_j = \min_i C_{ij} - u_i \quad \forall j \in J$. En effet, les contraintes de $(\mathcal{MK}_d)$ sont

$$u_i + v_j \leq C_{ij} \quad \forall (i,j) \in I \times J$$

Cela implique que

$$v_j \leq \min_i C_{ij} - u_i$$

Or, la fonction objectif de $(\mathcal{MK}_d)$ est $(u,v) \mapsto \sum_i u_i\mu_i + \sum_j v_j\nu_j$ avec $\nu_j \geq 0 \quad \forall j \in J$. Donc, pour maximiser la fonction-ojectif, on peut activer les contraintes $v_j \leq \min_i C_{ij} - u_i$, c'est-à-dire $v_j = \min_i C_{ij} - u_i$.

Cela nous permet d'exprimer $(\mathcal{MK}_d)$ comme un problème sans contraintes ne dépendant plus que de $\mathbf{u}$:

$$\max\left\{\sum_{i=1}^{N} \mu_i u_i + \sum_{j=1}^{N} \nu_j(\min_i C_{ij} - u_i) \mid \mathbf{u} \in \mathbb{R}^N\right\} \qquad (\mathcal{MK}_d)$$

Ensuite, en remarquant que $\max(z) = -\min(-z)$, on peut utiliser le résultat de convergence de f_ε précédemment démontré :

$$\min_i C_{ij} - u_i = -\max_i u_i - C_{ij} = \lim_{\varepsilon \to 0} -\varepsilon \log \left(\sum_{i=1}^{N} \exp\left(-\frac{C_{ij} - u_i}{\varepsilon} \right) \right)$$

Si l'on substitue cette expression dans la fonction-objectif de $(\mathcal{MK}_d)$ et on relaxe la limite, on obtient $\forall \varepsilon > 0$ le nouvel problème régularisé

$$\max \left\{ \sum_i u_i \mu_i - \sum_j \varepsilon \log \left(\sum_{i=1}^{N} \exp\left(-\frac{C_{ij} - u_i}{\varepsilon} \right) \right) \nu_j |\ \mathbf{u} \in \mathbb{R}^N \right\} \qquad (\mathcal{MK}_\varepsilon)$$

Ce problème est par construction sans contraintes, et on retrouve le problème $(\mathcal{MK}_d)$ en prenant la limite $\varepsilon \to 0$.

4.3 Mise en application sur l'exemple défini dans la partie 3.3.1

On reprend les données de la partie 3.3.1.

Un calcul direct nous donne que la k-ème composante du gradient du critère est

$$(\nabla \text{ critère })_k = \mu_k - \sum_j \nu_j \frac{\exp\left(-\frac{(C_{kj} - u_k)}{\varepsilon} \right)}{\sum_{i=1}^{N} \exp\left(-\frac{C_{ij} - u_i}{\varepsilon} \right)}$$

Ce gradient nous permet d'implémenter l'algorithme de gradient à pas fixe.

Voici les résultats obtenus pour $\varepsilon \in \{1, 0.5, 0.1, 0.05, 0.01, 0.005\}$:

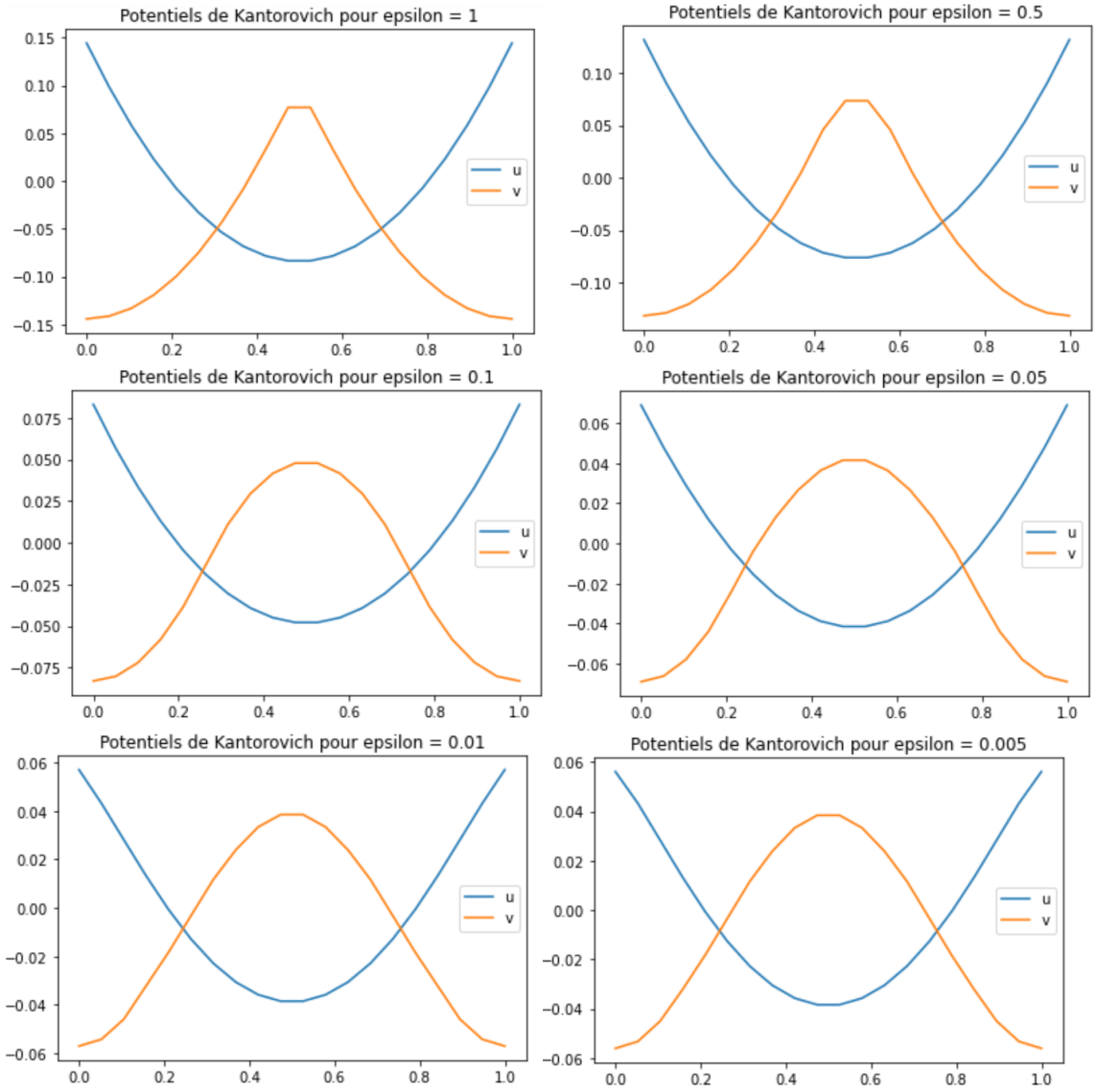


FIGURE 11

On veut maintenant illustrer la convergence de

$$\max\left\{\sum_i u_i\mu_i - \sum_j \varepsilon \log\left(\sum_{i=1}^{N} \exp\left(-\frac{C_{ij}-u_i}{\varepsilon}\right)\right)\nu_j \,|\, \mathbf{u}\in\mathbb{R}^N\right\} \qquad (\mathcal{MK}_\varepsilon)$$

vers

$$\max\left\{\sum_{i=1}^{N}\mu_i u_i+\sum_{j=1}^{N}\nu_j(\min_i C_{ij}-u_i)\mid \mathbf{u}\in\mathbb{R}^N\right\} \qquad (\mathcal{MK}_d)$$

On va donc calculer les valeurs des fonctions-objectif de $(\mathcal{MK}_\varepsilon)$ en les potentiels de Kantorovich pour $\varepsilon \in \{1, 0.5, 0.1, 0.05, 0.01, 0.005, 0.0003\}$:

```
epsilon :          Valeur de la fonction-objectif :
1.0000                  -2.867639163330509
0.5000                  -1.3774972796896876
0.1000                  -0.21944067973556636
0.0500                  -0.09035128403699619
0.0100                  -0.002312469012201748
0.0050                  0.005778293176521116
0.0010                  0.010417406173919617
0.0005                  0.0107712558893726
0.0003                  0.010907803468058867
```

On remarque que le nombre d'itérations nécessaires dans l'algorithme de gradient diminue lorsque ε tend vers 0 :

```
Le nombre d'itérations est  24106  pour epsilon =  1
Le nombre d'itérations est  12688  pour epsilon =  0.5
Le nombre d'itérations est  3094  pour epsilon =  0.1
Le nombre d'itérations est  2025  pour epsilon =  0.05
Le nombre d'itérations est  1341  pour epsilon =  0.01
Le nombre d'itérations est  1272  pour epsilon =  0.005
Le nombre d'itérations est  1072  pour epsilon =  0.001
Le nombre d'itérations est  727  pour epsilon =  0.0005
Le nombre d'itérations est  557  pour epsilon =  0.0003
```

On explique cela par le fait que l'algorithme de gradient se comporte mieux avec le gradient de la fonction-objectif de $(\mathcal{MK}_\varepsilon)$ lorsque ε tend vers 0.

On compare maintenant les potentiels de Kantorovich calculés par l'algorithme du simplex (voir ici) avec ceux calculés par l'algorithme de gradient avec ε petit ($\varepsilon = 0.005$).

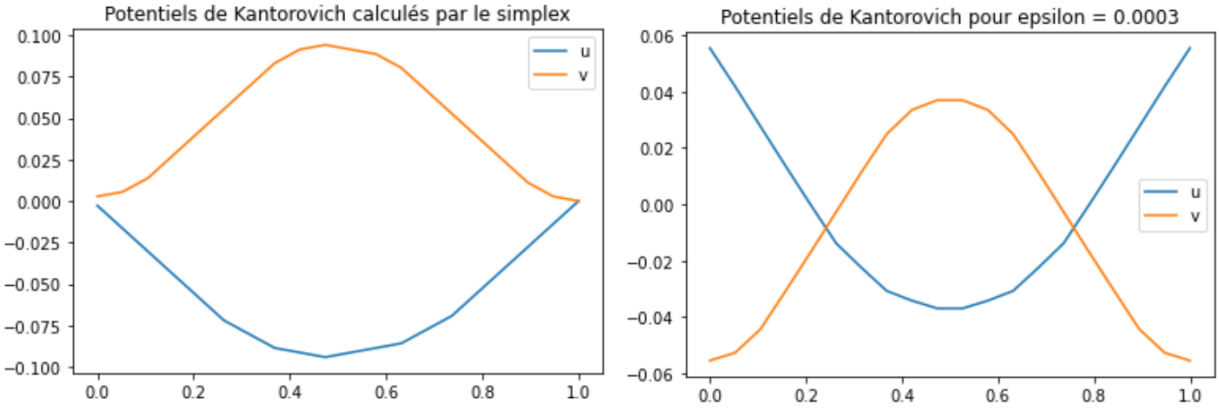


FIGURE 12

On remarque plusieurs choses :

— L'allure de **u** et **v** sont les mêmes ainsi que leur amplitude : un peu moins de 0.1.

— Les valeurs de **u** et **v** sont différentes. Dans le cas du simplex, $\mathbf{u} < 0$, $\mathbf{v} > 0$ et $\mathbf{u} < \mathbf{v}$, alors que dans le cas de l'algorithme de gradient, **u** et **v** changent de signe et de relation de supériorité-infériorité.

On peut expliquer l'aspect particulier des potentiels de Kantorovich dans le cas de l'algorithme de gradient. En effet, en regardant la figure 1 et la fonction-objectif de $(\mathcal{MK}_d)$, on peut prévoir la chose suivante : lorsque $\mu_i > \nu_i$, la composante u_i sera plus grande que la composante v_i pour maximiser la fonction-objectif (et inversement). Pour illustrer cet argument, on superpose la figure 1 et les potentiels de Kantorovich :

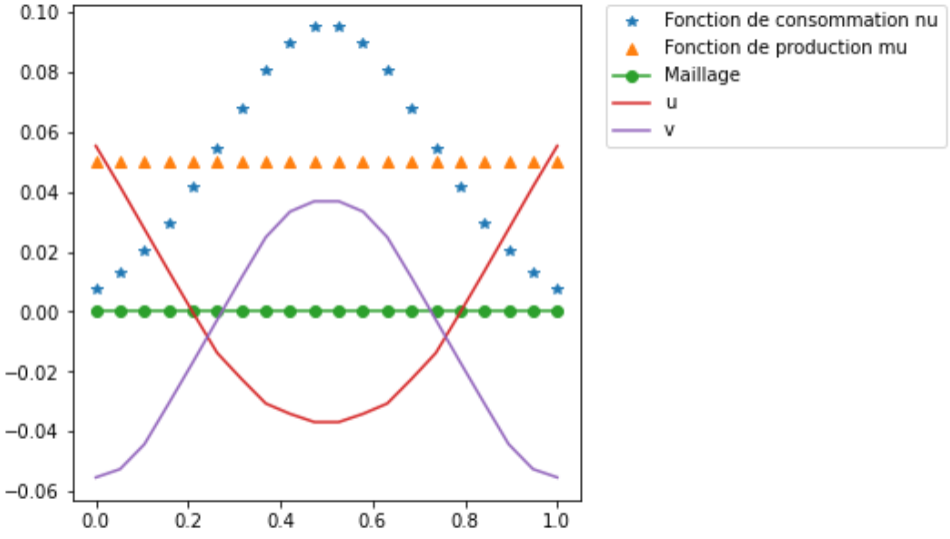


FIGURE 13

On voit graphiquement le fait que $\mu_i > \nu_i \Longrightarrow v_i > u_i$ et on le vérifie à la main dans le code.

On peut donc intuitivement se dire que les potentiels de Kantorovich calculés par l'algorithme de gradient sont "meilleurs" car ils profitent de la relation de supériorité-infériorité entre la fonction de production et celle de consommation. Cependant, on ne peut pas conclure que le résultat de l'algorithme de gradient est meilleur que celui du simplex. En effet, ce n'est qu'une intuition sur pourquoi les potentiels de Kantorovich ont cet aspect particulier et les valeurs de la fonction-objectif sont dans les deux cas similaires :

```
Valeur de la fonction-objectif de (MK_d) en les potentiels de Kantorovich calculés par
- le simplex :  0.011112315676793912
- l'algorithme de gradient :  0.011053381018070867
Erreur :  5.893465872304493e-05
```

De plus, les valeurs prises par les potentiels de Kantorovich dans le cas du simplex sont plus extrêmes : presque ± 0.1 contre ± 0.06. Ainsi, les contributions $\mu_i u_i$ et $\nu_i v_i$ dans la fonction-objectif sont plus importantes.

On avait résumer la stratégie des potentiels de Kantorovich dans le cas du simplex par "miser sur la forte consommation au centre". Dans le cas de l'algorithme de gradient, on peut la résumer par "miser sur la fonction la plus grande localement entre celle de production et celle de consommation".

5 La régularisation entropique

Dans cette troisième partie, on souhaite appliquer la régularisation entropique au problème primal $(\mathcal{MK})$. Elle a été introduite par Cuturi dans [7] et est très utilisée aujourd'hui dans des applications en machine learning. Elle permet de transformer le problème primal $(\mathcal{MK})$ en un problème sans contraintes $(\mathcal{H})$ avec de bonnes propriétés : convexité et une résolution moins coûteuse et plus simple lorsque les données sont nombreuses. On dualise ce problème dans l'optique d'utiliser l'algorithme de gradient à pas fixe comme dans la partie 4.3. Cependant, on se rend compte cette fois que cet algorithme ne donne pas satisfaction (instabilité de l'algorithme). On utilise donc un autre algorithme, appelé algorithme de Sinkhorn, qui repose sur les équations de Bernstein-Schrödinger. Ces dernières sont déduites des conditions d'optimalité et du lagrangien du problème $(\mathcal{H})$. Finalement, on compare les résultats obtenus avec ceux obtenus par le simplex dans la partie 3.3.1. On adopte donc la démarche suivante :

— Introduction du problème régularisé entropique $(\mathcal{H})$
— Détermination des équations de Bernstein-Schrödinger
— Etude du dual $\mathcal{H}_d$ du problème $\mathcal{H}$
— Mise en application de l'algorithme de gradient à pas fixe pour le problème $\mathcal{H}_d$
— Mise en application de l'algorithme de Sinkhorn pour le problème $\mathcal{H}_d$

5.1 Introduction du problème régularisé entropique

On reprend le problème primal $(\mathcal{MK})$ et on ajoute un terme d'entropie régularisé par ε à sa fonction-objectif. On obtient

$$F(\gamma) := \sum_{ij} C_{ij}\gamma_{ij} + \varepsilon \operatorname{Ent}(\gamma)$$

où $\operatorname{Ent} : \mathbb{R}_+^{N\times N} \to \mathbb{R}$ est définie par

$$\operatorname{Ent}(\gamma) = \sum_{ij} \gamma_{ij}\left(\log\left(\frac{\gamma_{ij}}{\mu_i\nu_j}\right) - 1\right)$$

Ce terme d'entropie pénalise la contrainte de positivité $\gamma \in \mathbb{R}_+^{N\times N}$ de $(\mathcal{MK})$. En effet, la fonction $\log$ ne prend que des termes strictement positifs.

On définit donc le problème entropique régularisé de $(\mathcal{MK})$:

$$\min\left\{F(\gamma) \;\middle|\; \sum_j \gamma_{ij} = \mu_i, \sum_i \gamma_{ij} = \nu_j\right\} \quad (\mathcal{H})$$

On observe immédiatement que lorsque $\varepsilon \to 0$, on retrouve le problème original.

Une première propriété qui montre que ce problème est intéressant est la convexité de la fonction-objectif F. Pour le prouver, on prend $\gamma, \beta \in \mathbb{R}^{N\times N}$ et on montre que $F\left(\gamma(1-t)+t\beta\right) \leqslant (1-t)F\left(\gamma\right) + tF\left(\beta\right)$.

On a

$$F\left(\gamma(1-t)+t\beta\right) = \sum_{i,j} C_{ij}\left(\gamma_{ij}(1-t)+\beta_{ij}t\right) + \varepsilon\left(\gamma_{ij}(1-t)+\beta_{ij}t\right)\left(\log\left(\frac{\gamma_{ij}(1-t)+\beta_{ij}t}{\mu_i\nu_j}\right) - 1\right)$$

Or, la fonction $g : x \mapsto C_{ij}x + \varepsilon x(\log(\frac{x}{\mu_i\nu_j}) - 1)$ est convexe comme somme de 2 fonctions convexes. La fonction $x \mapsto C_{ij}x$ est convexe car linéaire. La fonction $x \mapsto \varepsilon x(\log(\frac{x}{\mu_i\nu_j}) - 1)$ est convexe car elle est différentiable et $\frac{d^2}{dx^2}\varepsilon x(\log\frac{x}{a} - 1) = \frac{\varepsilon}{x}$ est strictement positive lorsque $x > 0$.

On a donc

$$\begin{aligned}
(1) &= \sum_{ij} g(\gamma_{ij}(1-t) + t\beta_{ij}) \\
&\leq \sum_{ij} tg(\beta_{ij}) + (1-t)g(\gamma_{ij}) \\
&= t\sum_{ij} g(\beta_{ij}) + (1-t)\sum_{ij} g(\gamma_{ij}) \\
&= tF(\beta) + (1-t)F(\gamma)
\end{aligned}$$

F est donc bien convexe.

On remarque également que l'ensemble admissible des contraintes d'égalité $X_E = \{\gamma \in \mathbb{R}_+^{N\times N} | \sum_j \gamma_{ij} = \mu_i, \sum_i \gamma_{ij} = \nu_j\}$ est lui-aussi convexe. En effet, soient $\gamma, \beta \in X_E$ et $t \in [0,1]$. Alors

$$\begin{aligned}
&\sum_j \gamma_{ij} t + \beta_{ij}(1-t) = t\mu_i + (1-t)\mu_i = \mu_i \\
&\sum_i \gamma_{ij} t + \beta_{ij}(1-t) = t\nu_j + (1-t)\nu_j = \nu_i \\
&\gamma t + \beta(1-t) \in \mathbb{R}_+^{N\times N}
\end{aligned}$$

donc $\gamma t + \beta(1-t) \in X_E$ et X_E est bien convexe.

Le problème entropique régularisé est donc convexe. Cette propriété est très puissante et nous fournit notamment des conditions d'optimalité que l'on va utiliser dans la partie suivante pour déterminer les équations de Bernstein-Schrödinger.

On écrit maintenant F sous une forme alternative. Par définition,

$$F(\gamma) = \sum_{ij} C_{ij}\gamma_{ij} + \varepsilon \sum_{ij} \gamma_{ij} \left(\log\left(\frac{\gamma_{ij}}{\mu_i \nu_j}\right) - 1 \right)$$

On peut mettre ε comme facteur, de manière à obtenir

$$F(\gamma) = \varepsilon \sum_{ij} \gamma_{ij} \left[-\left(-\frac{C_{ij}}{\varepsilon}\right) + \log\left(\frac{\gamma_{ij}}{\mu_i \nu_j}\right) - 1 \right]$$

Une petite astuce nous permet d'écrire cela comme

$$F(\gamma) = \varepsilon \sum_{ij} \gamma_{ij} \left[\log\left(\frac{\gamma_{ij}}{\mu_i \nu_j}\right) - \log\left(\exp\left(-\frac{C_{ij}}{\varepsilon}\right)\right) - 1 \right]$$

et donc, si $\bar{\gamma}_{ij} = \mu_i \nu_j \exp\left(\frac{-C_{ij}}{\varepsilon}\right)$, on réécrit directement cette expression comme

$$F(\gamma) = \varepsilon \sum_{ij} \gamma_{ij} \left[\log\left(\gamma_{ij}\right) - \log\left(\bar{\gamma}_{ij}\right) - 1 \right]$$

C'est-à-dire

$$F(\gamma) = \varepsilon \sum_{ij} \gamma_{ij} \left[\log\left(\frac{\gamma_i}{\bar{\gamma}_{ij}}\right) - 1 \right]$$

5.2 Détermination des équations de Bernstein-Schrödinger

On cherche maintenant à déterminer les équations de Bernstein-Schrödinger à partir des conditions d'optimalité et du lagrangien de $(\mathcal{H})$. Elles nous seront utiles pour dualiser $(\mathcal{H})$ et créer l'algorithme de Sinkhorn.

On commence par calculer le lagrangien. Pour un problème sous la forme

$$\begin{cases} \min F(\gamma) \\ c = 0 \end{cases}$$

le lagrangien est défini par

$$\ell(\gamma, \lambda) = F(\gamma) - \lambda \cdot C_E$$

où λ est le vecteur des multiplicateurs de Lagrange et C_E est le vecteur des contraintes d'égalité égal à

$$C_E = \begin{pmatrix} \sum_j \gamma_{1j} - \mu_1 \\ \sum_j \gamma_{2j} - \mu_2 \\ \vdots \\ \sum_j \gamma_{Nj} - \mu_N \\ \sum_i \gamma_{i1} - \nu_1 \\ \sum_i \gamma_{i2} - \nu_2 \\ \vdots \\ \sum_i \gamma_{iN} - \nu_N \end{pmatrix}$$

On a vu précédemment que le problème $(\mathcal{H})$ est convexe. De plus, la fonction-objectif F et la fonction définissant les contraintes d'égalité c_E (définie par le vecteur C_E) sont différentiables sur X_E. La dernière hypothèse de la condition suffisante d'ordre 1 d'un problème complexe est la condition de KKT, que l'on va maintenant étudier.

Calculons donc le gradient du lagrangien :

$$\nabla \ell = \nabla_\gamma F - \nabla_\gamma (\lambda \cdot C_E)$$

où

$$\begin{aligned} (\nabla_\gamma F)_{ij} &= \frac{\partial}{\partial \gamma_{ij}} F(\gamma) \\ &= \frac{\partial}{\partial \gamma_{ij}} \varepsilon \sum_{ij} \gamma_{ij} \left[\log\left(\frac{\gamma_i}{\bar{\gamma}_{ij}}\right) - 1 \right] \\ &= \varepsilon \frac{\partial}{\partial \gamma_{ij}} \left[\gamma_{ij} \left(\log\left(\frac{\gamma_{ij}}{\bar{\gamma}_{ij}}\right) - 1 \right) \right] \end{aligned}$$

Or, on a

$$\frac{d}{dx}\left[x \left(\log\left(\frac{x}{a}\right) - 1 \right) \right] = \left(\log\left(\frac{x}{a}\right) - 1 \right) + x \cdot \frac{a}{x} \cdot \frac{1}{a} = \log\left(\frac{x}{a}\right)$$

Donc

$$(\nabla_\gamma F)_{ij} = \varepsilon \log\left(\frac{\gamma_{ij}}{\bar{\gamma}_{ij}}\right)$$

D'autre part, on calcule les dérivées de $\lambda \cdot C_E$. Si l'on pose $\tilde{\lambda}_i = \lambda_{i+N}$, alors

$$\lambda \cdot C_E = \sum_{i=1}^{N} \lambda_i \left(\sum_{j=1}^{N} \gamma_{ij} - \mu_i\right) + \tilde{\lambda}_i \left(\sum_{j=1}^{N} \gamma_{ji} - \nu_i\right)$$

et donc

$$\frac{\partial}{\partial \gamma_{ij}}(\lambda \cdot C_E) = \lambda_i + \tilde{\lambda}_j$$

La condition de KKT nous donne donc les équations suivantes :

$$\varepsilon \log\left(\frac{\gamma_{ij}}{\bar{\gamma}_{ij}}\right) - \lambda_i - \tilde{\lambda}_j = 0$$

Ainsi, si $(\gamma^\star, (u, v))$ est une solution, alors il vérifie

$$\varepsilon \log\left(\frac{\gamma^*_{ij}}{\bar{\gamma}_{ij}}\right) - u_i - v_j = 0$$

Donc

$$\log\left(\frac{\gamma^*_{ij}}{\bar{\gamma}_{ij}}\right) = \frac{u_i}{\varepsilon} + \frac{v_j}{\varepsilon}$$

et alors

$$\gamma^*_{ij} = \bar{\gamma}_{ij} \exp\left(\frac{u_i}{\varepsilon}\right) \exp\left(\frac{v_j}{\varepsilon}\right)$$

C'est-à-dire

$$\gamma^*_{ij} = a_i b_j \bar{\gamma}_{ij}$$

où $a_i = \exp(u_i/\varepsilon)$, $b = \exp(v_j/\varepsilon)$, et u_i , v_j sont les multiplicateurs de Lagrange.

Si l'on somme pour $j \in J$, on obtient

$$\sum_j a_i b_j \bar{\gamma}_{ij} = \sum_j \gamma^*_{ij} = \mu_i \iff a_i = \frac{\mu_i}{\sum_j b_j \bar{\gamma}_{ij}}$$

par l'équation des contraintes et par la décomposition précédente de la solution γ^*. De même, si on somme pour $i \in I$, on a

$$\sum_i a_i b_j \bar{\gamma}_{ij} = \sum_i \gamma^*_{ij} = \nu_j \iff b_j = \frac{\nu_j}{\sum_i a_i \bar{\gamma}_{ij}}$$

Ces 2 équations sont les équations de Bernstein - Schrödinger.

5.3 Étude du dual

On va maintenant définir le problème dual $(\mathcal{H}_d)$ de $(\mathcal{H})$. Il est donné par

$$\sup_{u,v} \inf_{\gamma} \ell(\gamma, u, v)$$

Cependant, on sait que

$$\operatorname{argmin}_{\gamma} \ell(\gamma, u, v) = \gamma^*$$

C'est à dire, le problème dual est

$$\sup_{u,v} \ell(\gamma^*, u, v)$$

Or,

$$\ell\left(\gamma^*, u, v\right) = F\left(\gamma^*\right) + \sum_i u_i (C_E)_i + \sum_j v_j (C_E)_j$$

où

$$\begin{aligned} F\left(\gamma^*\right) &= \varepsilon \sum_{ij} a_i b_i \bar{\gamma}_{ij} \left(\log\left(a_i\right) + \log\left(b_j\right) - 1\right) \\ &= \varepsilon \sum_{ij} a_i b_j \bar{\gamma}_{ij} \left(\frac{u_i}{\varepsilon} + \frac{v_j}{\varepsilon} - 1\right) \end{aligned}$$

Alors

$$\begin{aligned} \ell(\gamma^*, u, v) &= \sum_{ij} a_i b_j \bar{\gamma}_{ij} \left(u_i + v_j - \varepsilon\right) + \sum_i u_i (C_E)_i + \sum_j v_j (C_E)_j \\ &= \sum_{i,j} a_i b_j \bar{\gamma}_{ij} u_i + \sum_{i,j} a_i b_j \bar{\gamma}_{ij} v_j - \varepsilon \sum_{ij} a_i b_j \bar{\gamma}_{ij} + \sum_i u_i (C_E)_i + \sum_j v_j (C_E)_j \\ &= \sum_i \mu_i u_i + \sum_j \nu_j v_j - \varepsilon \sum_{ij} \exp\left(\frac{u_i + v_j - C_{ij}}{\varepsilon}\right) + \sum_i u_i \left(\sum_j \gamma^*_{ij} - \mu_i\right) + \sum_j v_j \left(\sum_i \gamma^*_{ij} - \nu_j\right) \\ &= \sum_{ij} u_i \gamma^*_{ij} + \sum_{ij} v_j \gamma^*_{ij} - \varepsilon \sum_{ij} \exp\left(\frac{u_i + v_j - C_{ij}}{\varepsilon}\right) \\ &= \sum_i u_i \mu_i + \sum_j v_j \nu_j - \varepsilon \sum_{ij} \exp\left(\frac{u_i + v_j - C_{ij}}{\varepsilon}\right) \text{ où on a utilisé } \sum_j \gamma^\star_{ij} = \mu_i \text{ et } \sum_i \gamma^\star_{ij} = \nu_j \end{aligned}$$

Donc le problème dual s'écrit

$$\sup_{u,v} \left\{ \sum_i u_i \mu_i + \sum_j v_j \nu_j - \varepsilon \sum_{ij} \exp\left(\frac{u_i + v_j - C_{ij}}{\varepsilon}\right) \right\} \qquad (\mathcal{H}_d)$$

qui est un problème sans contraintes.

Tout comme le problème régularisé de la partie 4.2, on se ramène au problème dual $(\mathcal{MK}_d)$ en prenant $\varepsilon \to 0$. En effet, on a

$$\lim_{\varepsilon \to 0} \varepsilon \sum_{ij} \exp(\frac{u_i + v_j - C_{ij}}{\varepsilon}) = 0$$

car $u_i + v_j - C_{ij} \leq 0$ d'après les contraintes de $(\mathcal{MK}_d)$. Donc, lorsque $\varepsilon \to 0$, les fonctions-objectifs de $(\mathcal{H}_d)$ et $(\mathcal{MK}_d)$ sont les mêmes.

On souhaite donc résoudre $(\mathcal{H}_d)$ lorsque ε est petit, avec un algorithme plus simple et moins coûteux que celui du simplex. Comme dans la partie 4.3, on va essayer d'utiliser celui de gradient à pas fixe.

5.4 Mise en application sur l'exemple défini dans la partie 3.3.1

Le problème $(\mathcal{H}_d)$ est équivalent à

$$-\inf_{u,v} \left\{ -\sum_i u_i \mu_i - \sum_j v_j \nu_j + \varepsilon \sum_{ij} \exp\left(\frac{u_i + v_j - C_{ij}}{\varepsilon}\right) \right\}$$

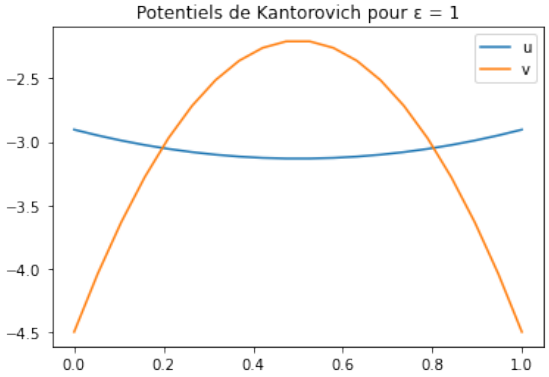
On va utiliser un algorithme de gradient sur ce problème. On peut le faire car la fonction-objectif est différentiable. De plus, on peut remarquer qu'elle est convexe (grâce à la convexité des fonctions linéaire et exponentielle), donc les points stationnaires trouvés par l'algorithme de descente de gradient sont des minimums globaux.

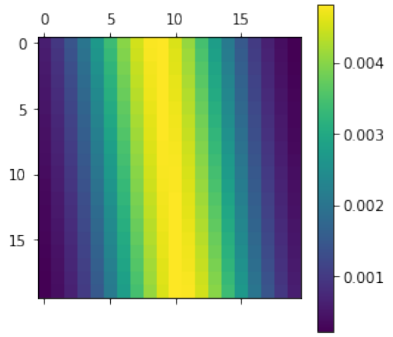
On calcule le gradient du critère :

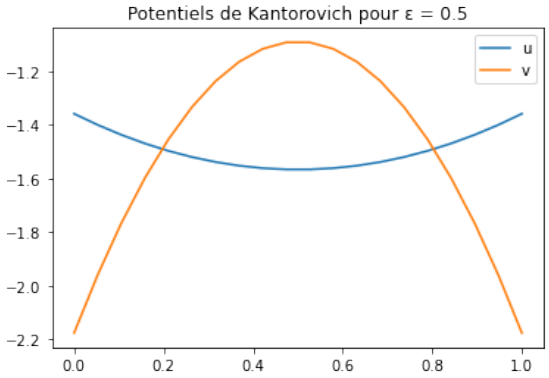
$$\frac{\partial \text{critère}}{\partial u_i} = -\mu_i + \sum_j \exp\left(\frac{u_i + v_j - C_{ij}}{\varepsilon}\right) = -\mu_i + \sum_j a_i b_j \bar{\gamma}_{ij}$$

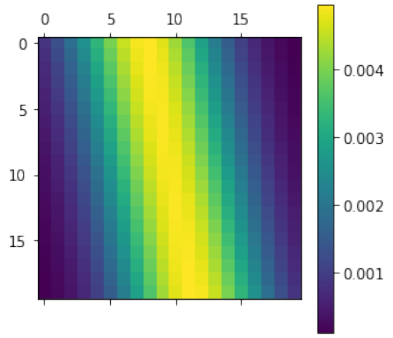
$$\frac{\partial \text{critère}}{\partial v_j} = -\nu_i + \sum_j \exp\left(\frac{u_i + v_j - C_{ij}}{\varepsilon}\right) = -\nu_i + \sum_j a_i b_j \bar{\gamma}_{ij}$$

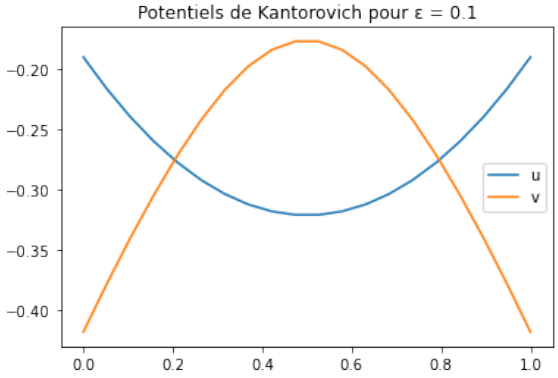
Une fois que l'on a trouvé des potentiels u et v, on peut retrouver γ^* en utilisant l'équation $\gamma^*_{ij} = a_i b_j \bar{\gamma}_{ij}$. Voici les résultats pour différentes valeurs du paramètre de régularisation ε.

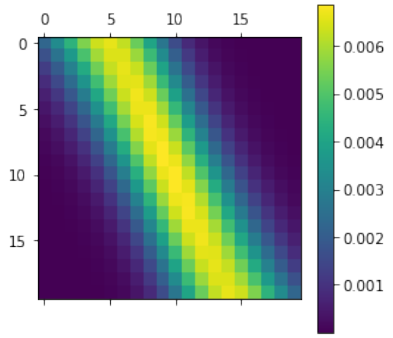

Potentiels de Kantorovich pour ε = 1
u
v
−2.5
−3.0
−3.5
−4.0
−4.5
0.0
0.2
0.4
0.6
0.8
1.0

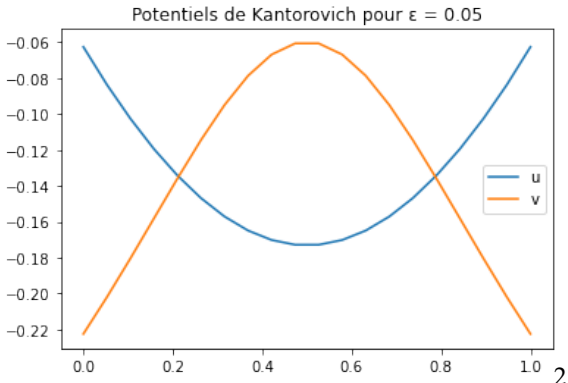

0
5
10
15
0.004
0.003
0.002
0.001

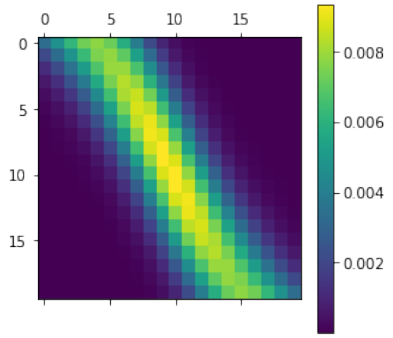

Potentiels de Kantorovich pour ε = 0.5
u
v
−1.2
−1.4
−1.6
−1.8
−2.0
−2.2
0.0
0.2
0.4
0.6
0.8
1.0


0
5
10
15
0.004
0.003
0.002
0.001


Potentiels de Kantorovich pour ε = 0.1
u
v
−0.20
−0.25
−0.30
−0.35
−0.40
0.0
0.2
0.4
0.6
0.8
1.0


0
5
10
15
0.006
0.005
0.004
0.003
0.002
0.001


Potentiels de Kantorovich pour ε = 0.05
u
v
−0.06
−0.08
−0.10
−0.12
−0.14
−0.16
−0.18
−0.20
−0.22
0.0
0.2
0.4
0.6
0.8
1.0


0
5
10
15
0.008
0.006
0.004
0.002

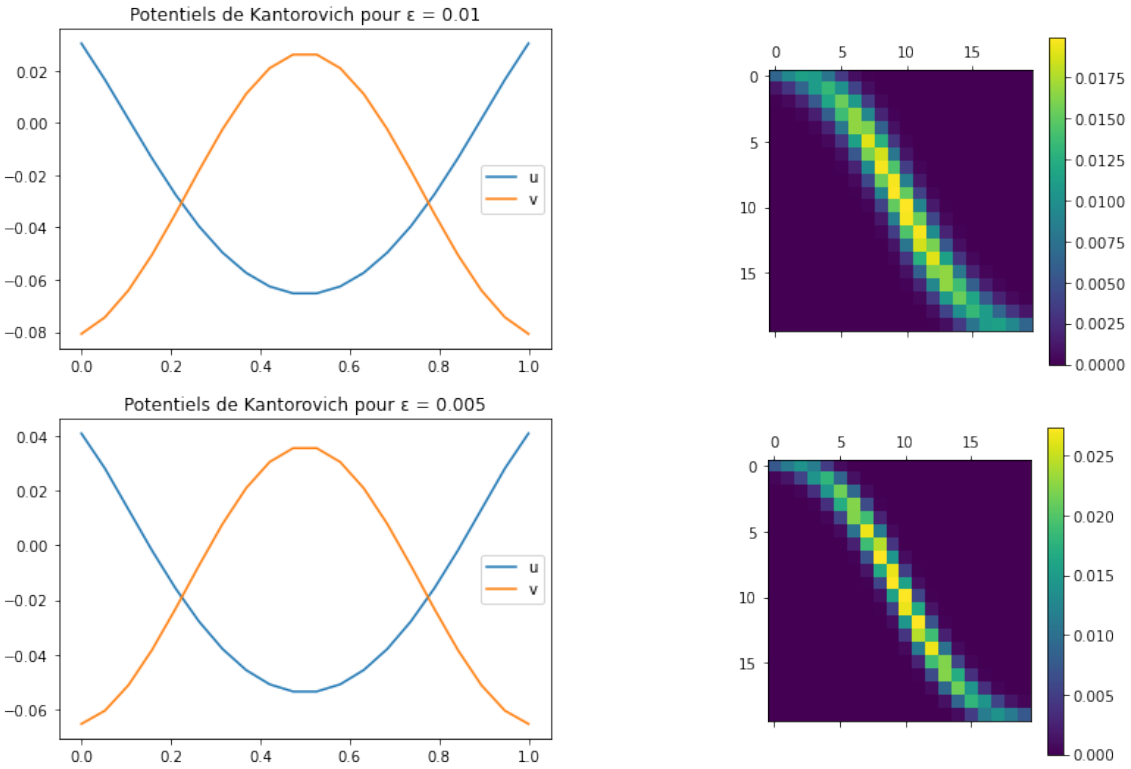
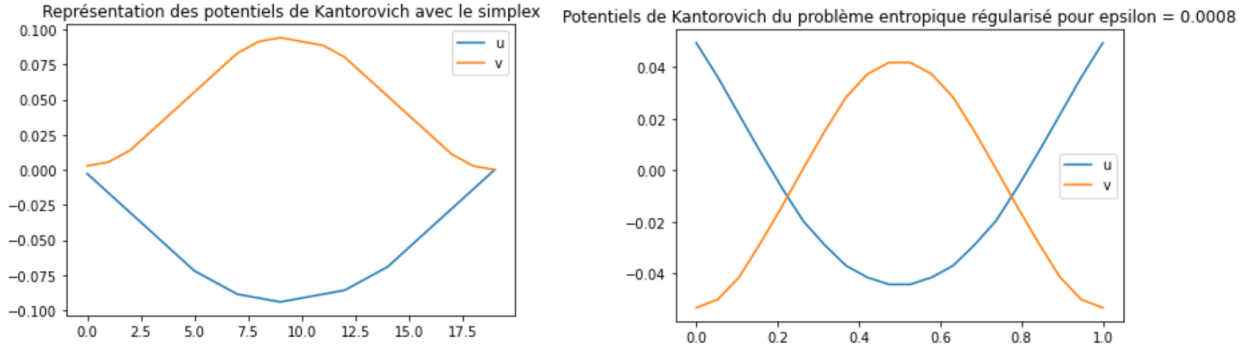


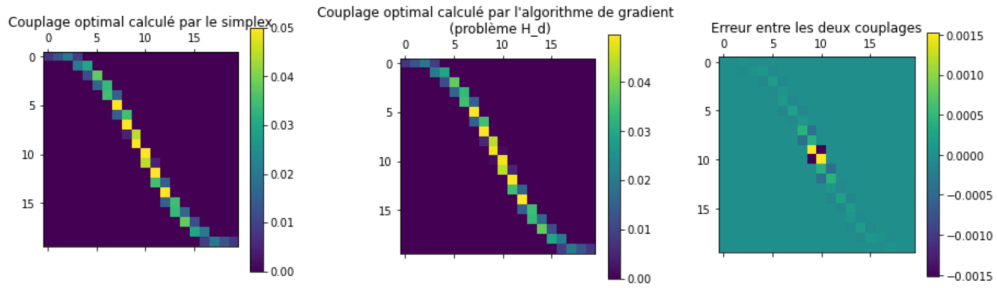
FIGURE 14

Comparons maintenant les potentiels de Kantorovich calculés par l'algorithme du simplex avec ceux calculés par l'algorithme de gradient avec ε petit ($\varepsilon = 0.0008$, la valeur la plus petite de ε pour laquelle l'algorithme fonctionne).



Visuellement, on vérifie que les potentiels à partir de la résolution de $\mathcal{H}_d$ par descente de gradient sont identiques à ceux obtenus dans la régularisation avec $log - sum - exp$. On peut ainsi appliquer la même analyse sur leur "stratégie" (voir ici).

Comparons maintenant le couplage calculé par l'algorithme du simplex avec celui calculé par l'algorithme de gradient (toujours avec $\varepsilon = 0.0008$).



On voit que les 2 couplages sont très similaires avec une erreur de l'ordre de 10^{-3}.

Maintenant on veut vérifier l'équivalence décrite ci-dessus. C'est-à-dire, on veut montrer que

$$\sup_{u,v} \left\{ \sum_i u_i \mu_i + \sum_j v_j v_j - \varepsilon \sum_{ij} \exp\left(\frac{u_i + v_j - C_{ij}}{\varepsilon} \right) \right\} \quad (\mathcal{H}_d)$$

avec les résultats de cette descente de gradient et lorsque $\varepsilon \to 0$ est équivalent à

$$\max \left\{ \sum_{i=1}^{N} \mu_i u_i + \sum_{j=1}^{N} v_j \left(\min_i C_{ij} - u_i \right) \mid \mathbf{u} \in \mathbb{R}^N \right\} \quad (\mathcal{MK}_d)$$

Or, ce dernier problème a déjà été résolu par le simplex et par la descente de gradient de la régularisation $log - sum - exp$. Donc pour conclure l'équivalence, on compare la fonction-objectif obtenue avec ces trois méthodes :

```
Valeur de la fonction-objectif de (MK_d) en les potentiels de Kantorovich calculés par
- le simplex :  0.011112315676793914

- l'algorithme de gradient pour (H_d) :  0.00833302463420467
Erreur :  0.0027792910425892443

- l'algorithme de gradient pour (MK_epsilon) :  0.011053381018070867
Erreur :  5.8934658723046665e-05
```

On remarque que le résultat le moins bon, c'est-à-dire la valeur de la fonction-objectif la plus basse dans le problème de maximisation $(\mathcal{MK}_d)$, est celui du problème entropique régularisé $(\mathcal{H}_d)$ avec les potentiels calculés par l'algorithme de gradient. De plus, l'erreur par rapport au résultat du simplex est bien plus importante que celle du problème $(\mathcal{MK}_\varepsilon)$. La première est de l'ordre 10^{-3} et la seconde d'ordre 10^{-5}.

De plus, on remarque que l'algorithme de gradient pour calculer $(\mathcal{H}_d)$ est instable pour des petites valeurs de ε. C'est pourquoi, on cherche un autre algorithme pour résoudre $(\mathcal{H}_d)$. Cet algorithme est celui de Sinkhorn qui se base sur les équations de Berntein-Schrödinger.

5.5 L'algorithme de Sinkhorn

On a vu que la solution du problème $(\mathcal{MK})$ peut s'écrire comme

$$\gamma^*_{ij} = a_i b_j \bar{\gamma}_{ij}$$

où la matrice $\bar{\gamma}$ peut être construite directement des mesures μ et ν. Donc, si l'on trouve une façon de construire les coefficients a_i et b_j sans résoudre directement le problème dual, on aura une manière alternative de résoudre le problème $(\mathcal{MK})$ qui est potentiellement plus rapide que la descente de gradient pour le problème dual.

On observe les équations de Bernstein-Schrödinger :

$$a_i = \frac{\mu_i}{\sum_j b_j \bar{\gamma}_{ij}} \quad b_j = \frac{\nu_j}{\sum_i a_i \bar{\gamma}_{ij}}$$

Les 2 équations sont dépendantes. Cela suggère d'écrire un algorithme itérative de la forme

$$a_i^{n+1} = \frac{\mu_i}{\sum_j b_j^n \bar{\gamma}_{ij}} \quad b_j^{n+1} = \frac{\nu_j}{\sum_i a_i^{n+1} \bar{\gamma}_{ij}}$$

L'algorithme est appelé Sinkhorn car Richard Sinkhorn et Paul Knopp ont prouvé sa convergence [8]. On l'initialise par

$$a_i^0 = \exp\left(\frac{u_i^0}{\varepsilon}\right)$$

$$b_j^0 = \exp\left(\frac{v_j^0}{\varepsilon}\right)$$

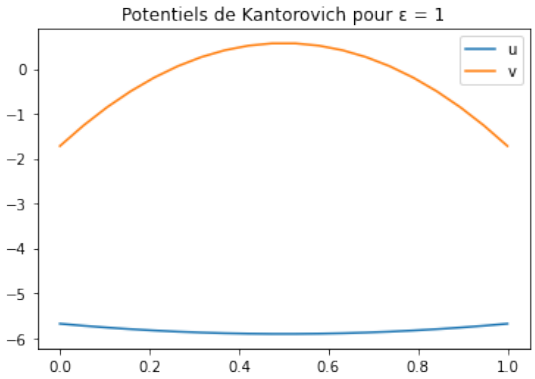
Une fois que l'algorithme a convergé, on peut construire les potentiels **u** et **v** avec

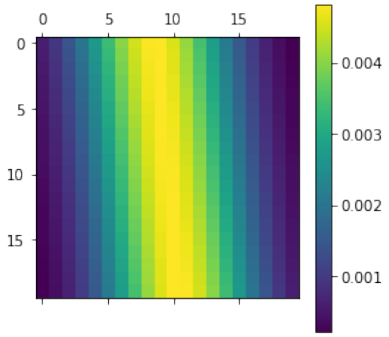
$$a = \exp\left(\frac{u}{\varepsilon}\right) \quad b = \exp\left(\frac{v}{\varepsilon}\right)$$

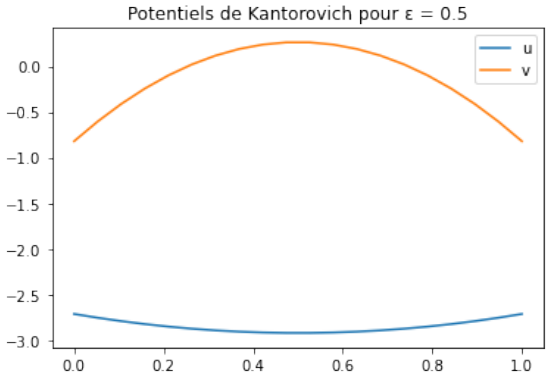
C'est-à-dire,

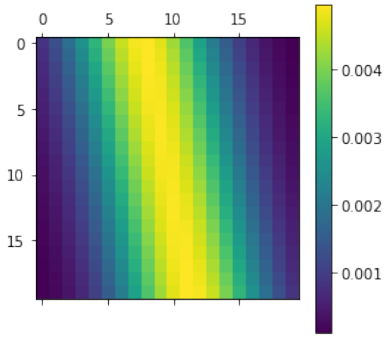
$$u = \varepsilon \log(a) \quad v = \varepsilon \log(b)$$

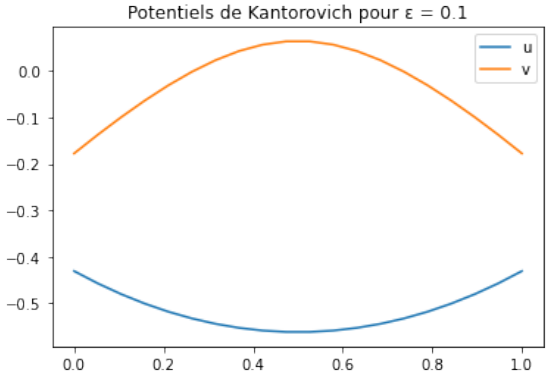
Voici les résultats que l'on a obtenu en utilisant l'algorithme de Sinkhorn, avec $\varepsilon \in \{1, 0.5, 0.1, 0.05, 0.01, 0.005\}$.

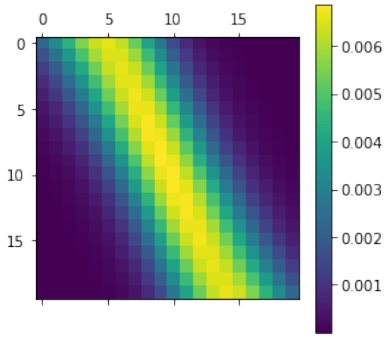

Potentiels de Kantorovich pour ε = 1
u
v

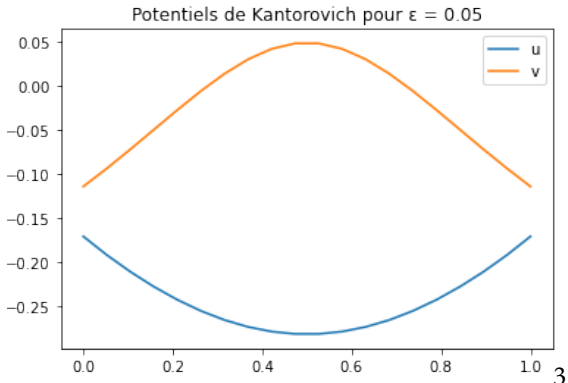


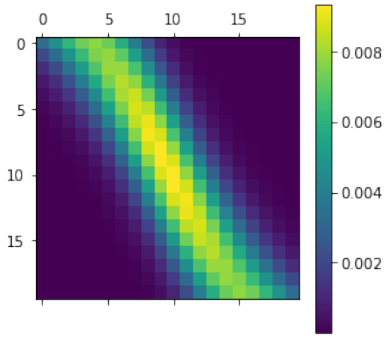

Potentiels de Kantorovich pour ε = 0.5
u
v




Potentiels de Kantorovich pour ε = 0.1
u
v




Potentiels de Kantorovich pour ε = 0.05
u
v



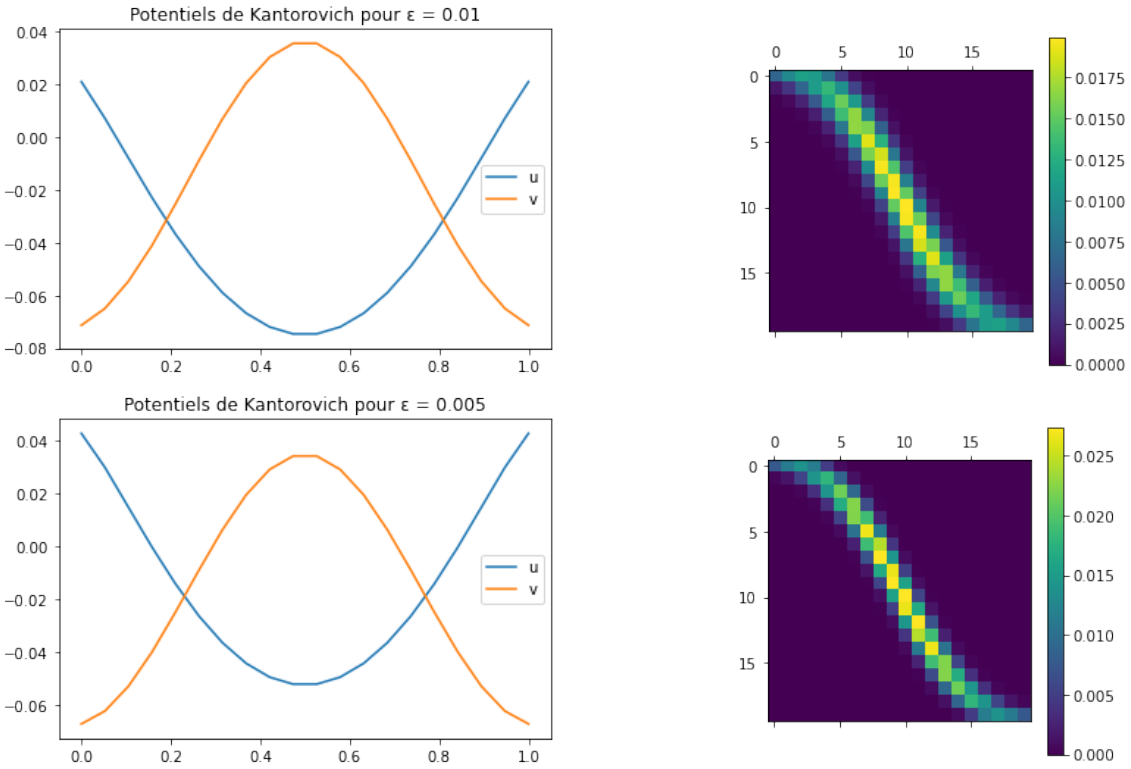
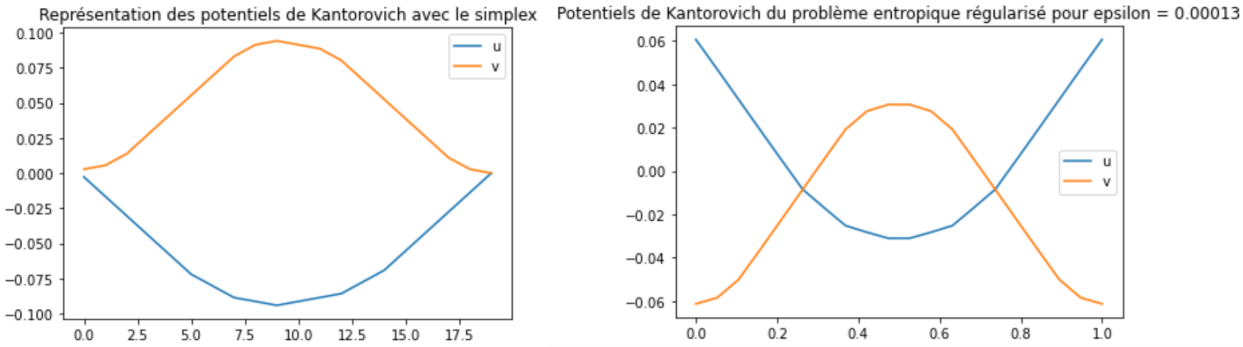


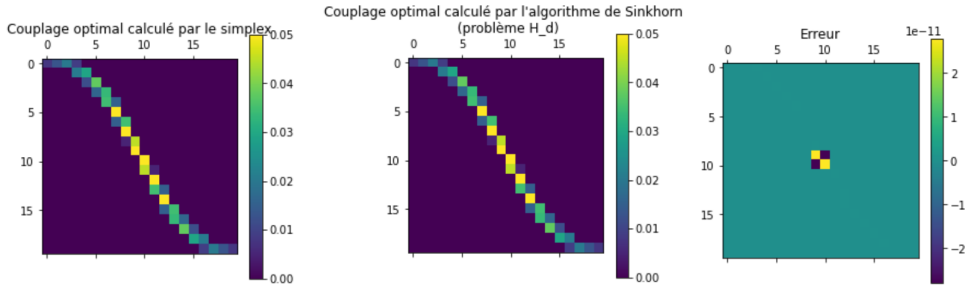
FIGURE 15

Comparons maintenant les potentiels de Kantorovich calculés par l'algorithme du simplex avec ceux calculés par l'algorithme de Sinkhorn avec ε petit ($\varepsilon = 0.00013$, la valeur la plus petite de ε pour laquelle l'algorithme fonctionne).



Comme avec l'algorithme de gradient, on voit que les potentiels sont identiques à ceux obtenus dans la régularisation avec $log - sum - exp$. On peut ici aussi appliquer la même analyse sur leur "stratégie" (voir ici).

Comparons maintenant le couplage calculé par l'algorithme du simplex avec celui calculé par l'algorithme de Sinkhorn (toujours avec $\varepsilon = 0.00013$).



On voit que les 2 couplages sont presque identiques avec seulement 4 valeurs différentes. Les erreurs sont de l'ordre de 10^{-11}, on peut donc considérer ces 2 couplages comme étant égaux.

Ensuite, on montre de nouveau l'équivalence entre $(\mathcal{MK}_d)$ et $(\mathcal{H}_d)$ lorsque $\varepsilon \to 0$.

```
Valeur de la fonction-objectif de (MK_d) en les potentiels de Kantorovich calculés par
- le simplex :  0.011112315676793912

- l'algorithme de gradient pour (H_d) :  0.008833024634204668
Erreur :  0.0027792910425892443

- l'algorithme de Sinkhorn pour (H_d) :  0.010661326787759502
Erreur :  0.0004509888890344097

- l'algorithme de gradient pour (MK_epsilon) :  0.011053381018070867
Erreur :  5.893465872304493e-05
```

On voit que l'algorithme de Sinkhorn donne un meilleur résultat par rapport à l'algorithme de gradient car $0.010661326787759502 > 0.008833024634204668$. De plus, l'erreur par rapport au résultat du simplex est de l'ordre de 10^{-4} et elle est plus faible que celle de l'algorithme de gradient (de l'ordre de 10^{-3}).

L'algorithme de Sinkhorn est donc meilleur que l'algorithme de gradient pour résoudre $(\mathcal{H}_d)$. D'autre part, la stabilité de l'algorithme de Sinkhorn est meilleure lorsque ε tend vers 0 que celle de l'algorithme de gradient. En effet, on a pu utiliser $\varepsilon = 0.00013$ contre $\varepsilon = 0.0008$.

6 Wasserstein flot pour le problème de matching

6.1 Un peu de géométrie

Une observation clé de la théorie est que la solution du problème de Monge-Kantorovich nous permet de définir une distance dans l'espace de mesures. Cette distance est appelé la distance de

Wasserstein, et est définie comme

$$\mathcal{W}_2^2(\mu,\nu) := (\mathcal{MK})$$

où $(\mathcal{MK})$ est la valeur de la fonction-objectif dans la solution du problème de Monge-Kantorovich pour les mesures μ et ν.

Quand on a une distance dans un espace, on peut définir les courbes qui minimisent localement cette distance. Ces courbes sont appelés géodésiques, et elles peuvent être vues aussi comme les courbes qui minimisent une fonctionnelle d'énergie. Un tel point de vue est parfois appelé une formulation variationnelle. Dans notre cas, avec mesures discrètes sur $\mathbb{R}^2$, si l'on fixe une mesure cible de transport $\nu = \sum_j \delta_{(\nu_j^x,\nu_j^y)}$, on définit la fonctionnelle d'énergie en fonction d'une mesure μ à transporter vers ν :

$$\mathcal{E}(\mu) := \mathcal{W}_\varepsilon\left(\frac{1}{n}\sum_i \delta_{(\mu_i^x,\mu_i^y)}, \frac{1}{m}\sum_j \delta_{(\nu_j^x,\nu_j^y)}\right)$$

où $(\nu_i^x,\nu_i^y) \in \mathbb{R}^2$ sont les coordonnées du i-ème point de ν (idem pour μ).

Donc le problème de trouver les géodésiques devient équivalent à résoudre le problème de Monge-Kantorovich, et en même temps équivalent à minimiser une fonctionnelle d'énergie. Si l'on arrive à calculer le gradient de l'énergie, on peut utiliser une descente de gradient pour trouver le chemin minimisant dans l'espace de mesures (qui avec cette distance est appelé espace de Wasserstein).

Comme avant, on peut aussi travailler avec la "distance" (et par conséquent l'énergie) régularisée :

$$\mathcal{W}_\varepsilon(\mu,\nu) := \sum_{ij} C_{ij}\gamma_{ij} + \varepsilon\sum_{ij}\gamma_{ij}\left(\log\left(\frac{\gamma_{ij}}{\mu_i\nu_j}\right) - 1\right)$$

6.2 Applications

Cherchons le gradient de l'énergie régularisée. Pour calculer le gradient de $W_\varepsilon(\mu,\nu)$, on utilise une idée présentée dans [3]. On observe que $W_\varepsilon(\mu,\nu)$ est la valeur de la fonction-objectif associée à la solution du problème de Monge-Kantorovich perturbé par ε. Le problème primal dépend seulement des poids des mesures μ et ν (et pas de ses positions), et de la matrice de coût C_{ij}. On écrit donc $W_\varepsilon(\mu,\nu) = L(\varepsilon, C_{ij}, \mu, \nu)$.

On peut introduire la dépendance en la position des poids μ_i dans le problème via C_{ij} . On notera $C_{ij}(z) = C_{ij}(\mu_1^x, \mu_2^x, \ldots, \mu_n^x, \mu_1^y, \ldots \mu_n^y\mu)$ pour ν fixée. Notre énergie s'écrit donc

$$\mathcal{E}(x) = L(\varepsilon, C_{ij}(x), \mu, \nu)$$

On assume les différentiabilités nécessaires (le terme entropique est important pour cela) [9]. On prend le gradient et on utilise la règle de la chaîne

$$\nabla \mathcal{E}(x) = \nabla L(C) \cdot \nabla C(x)$$

Or,

$$(\nabla L(C))_{ij} = \frac{\partial}{\partial C_{ij}} \sum_{ij} C_{ij}\gamma_{ij} + \varepsilon \sum_{ij} \gamma_{ij} \left(\log \left(\frac{\gamma_{ij}}{\mu_i \nu_j} \right) - 1 \right) = \gamma_{ij}$$

C'est-à-dire,

$$\nabla L(C) = \gamma$$

donc

$$\nabla \mathcal{E}(x) = \gamma \cdot \nabla C(x)$$

Pour le coût quadratique,

$$(\nabla C(x))_{\mu_i^x} = \frac{1}{2} \frac{\partial}{\partial \mu_i^x} |\mu_i^x - \nu_j^x|^2 + |\mu_i^y - \nu_j^y|^2 = |\mu_i^x - \nu_j^x| \operatorname{sgn}(\mu_i^x - \nu_j^x) = \mu_i^x - \nu_j^x$$

Donc

$$(\nabla \mathcal{E}(x))_i = \sum_j \gamma_{ij}(\mu_i^x - \nu_j^x) = \mu_i \mu_i^x - \sum_j \gamma_{ij} \nu_j^x$$

où on a utilisé $\sum_j \gamma_{ij} = \mu_i$. La valeur de γ peut être obtenue à chaque itération avec l'algorithme de Sinkhorn. On remarque que c'est aussi important de mettre à jour la matrice de coût à chaque itération, car les points de μ ont changé de position.

6.2.1 Mise en application sur l'exemple défini dans la partie 3.3.2

On peut tester Sinkhorn dans le cas 2D en prenant des distributions définies comme 3.3.2. C'est-à-dire, on prend 20 points avec distribution uniforme pour définir μ, et 20 points avec distribution normale centrée pour définir ν :

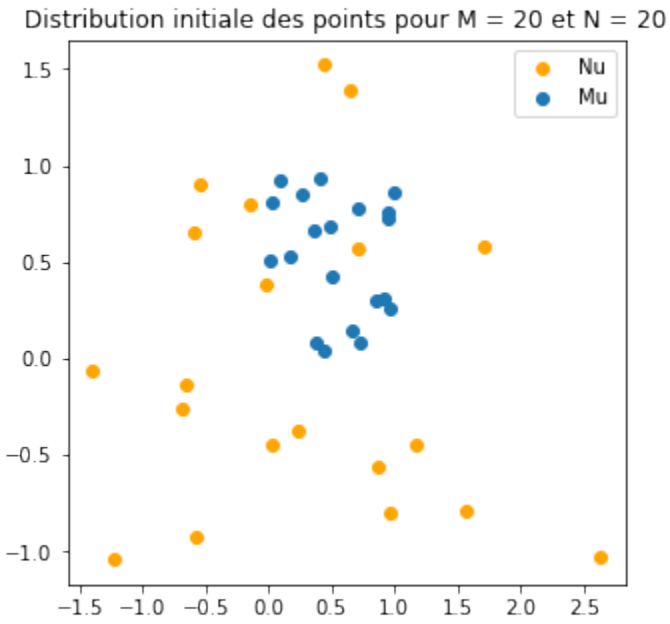


Figure 16

On obtient les potentiels et la matrice de transport suivants

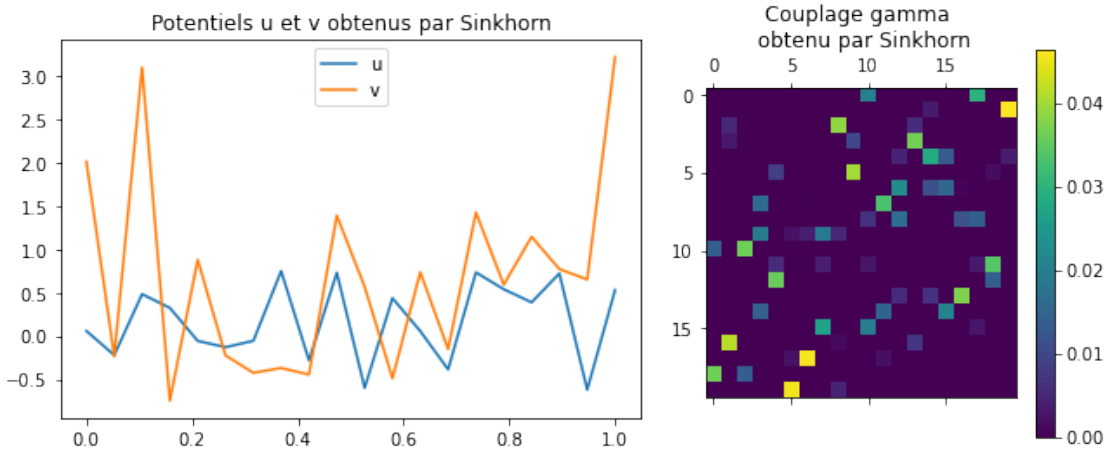


Figure 17

Comparons ce résultat avec le résultat de la partie 3.3.2 (résolution du primal avec le simplex) :

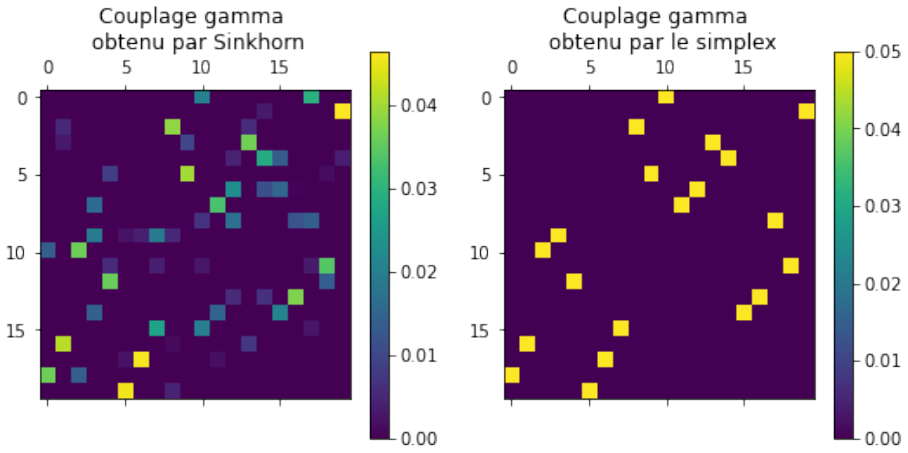


Figure 18

La matrice de couplage γ calculée par Sinkhorn semble correcte si on la compare à celle calculée par le simplex. En effet, les coefficients les plus grands semblent uniques sur leur ligne et leur colonne. De plus, ils coïncident avec les coefficients non nuls de la matrice de permutation. On s'attend donc à ce que le résultat de Sinkhorn converge vers celui du simplex lorsque $\varepsilon \to 0$.

On représente le champ du gradient à l'état initial :

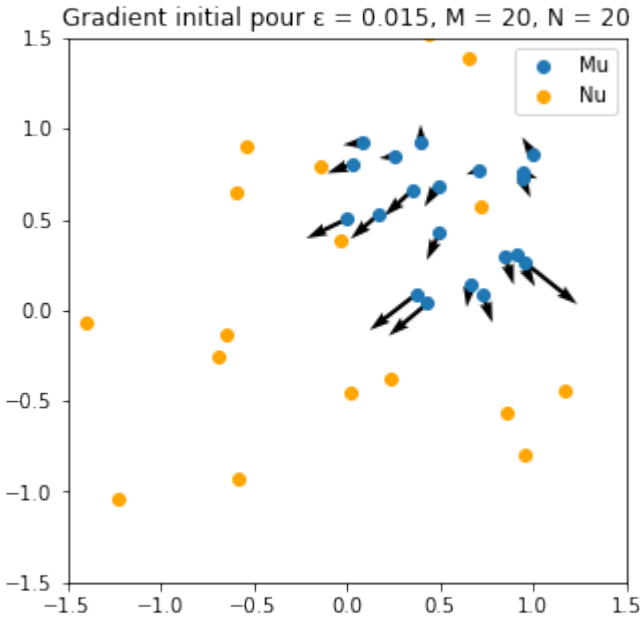


Figure 19

Le champ de gradient semble bien pointer de μ vers ν.

6.2.2 Mise en application sur un exemple en 2D

Passons maintenant au cas avec un nombre plus grand de points, distribués initialement selon

$$\mu^0 = \frac{1}{N}\sum_{i=1}^{N} \delta_{z_i^0}, \quad \nu^0 = \frac{1}{M}\sum_{i=1}^{M} \delta_{y_i^0}$$

où `z0 = np.random.rand(2,N)-.5` et `y = np.vstack((np.cos(theta)*r,np.sin(theta)*r))` avec `theta = 2*np.pi*np.random.rand(1,N)` et `r = .8 + .2*np.random.rand(1,N)`.
Les points z_i sont uniformément répartis dans le carré $[-0.5, 0.5] \times [-0.5, 0.5]$ et les y_i sont répartis sur le cercle $\mathcal{C}(0, 0.8)$ où on a jouté un bruit positif uniforme sur $[0, 0.2]$ pour le rayon.

Supposons que chaque mesure a 100 points. La distribution initiale de points est la suivante :

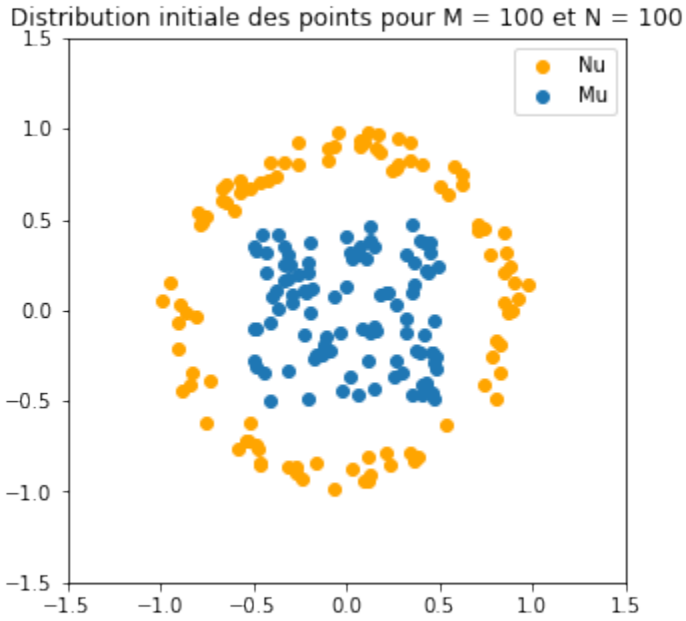


Figure 20

On représente le champ du gradient à l'état initial :

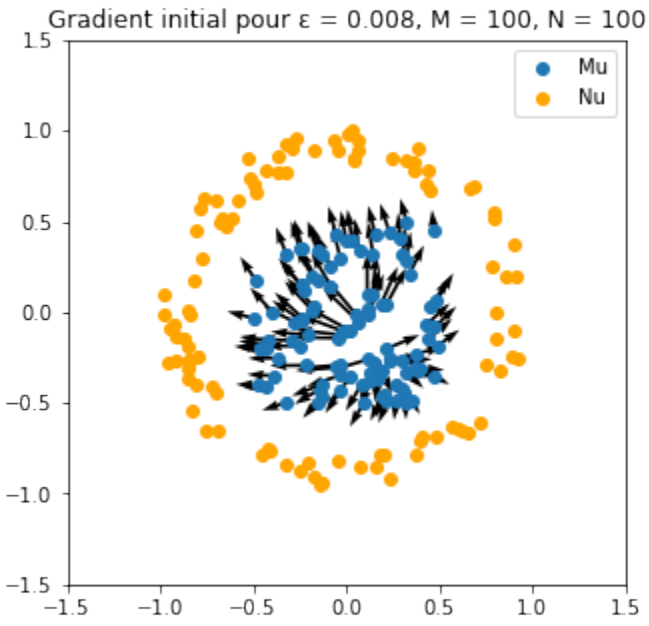


FIGURE 21

On fait une descente de gradient de pas $\tau = 0.1$, et on compare ce qu'on obtient pour des différents valeurs de ε. On prend $\varepsilon \in \{1, 0.5, 0.1, 0.05, 0.01, 0.008\}$. On graphe la distribution μ toutes les 5 itérations.

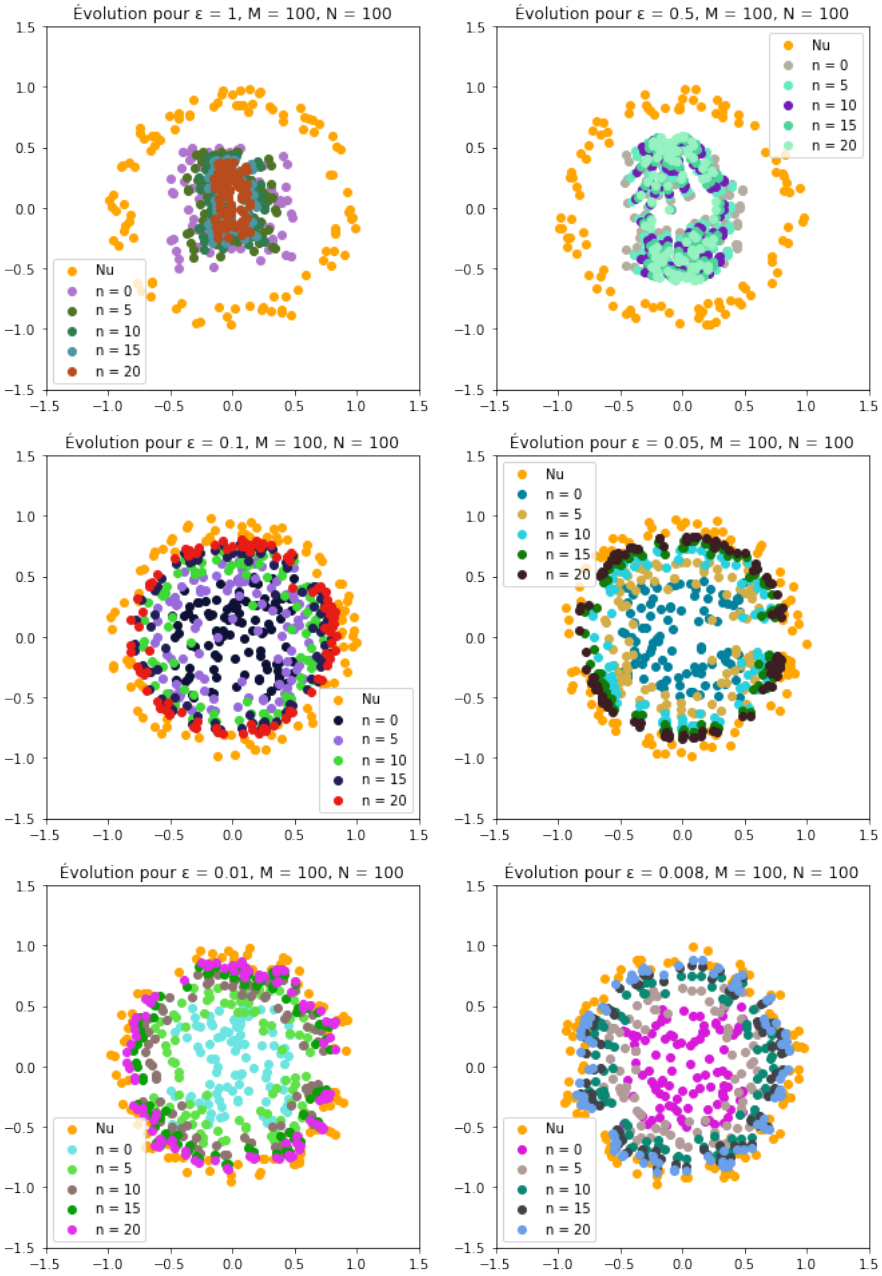


FIGURE 22

On observe que pour des grandes valeurs de ε (lorsque $\varepsilon \in \{1, 0.5\}$), la mesure μ se concentre proche du centre du cercle formé par les points de ν. Cependant, lorsque ε devient petit, on observe bien une convergence vers la mesure cible. De plus, on observe un phénomène de matching : chaque

point constituant la mesure ν voit converger vers lui un point constituant la mesure μ.

Reprenons cette expérience avec $M = 200$, c'est-à-dire qu'on double le nombre de points dans la mesure cible ν. On observe l'évolution suivante :

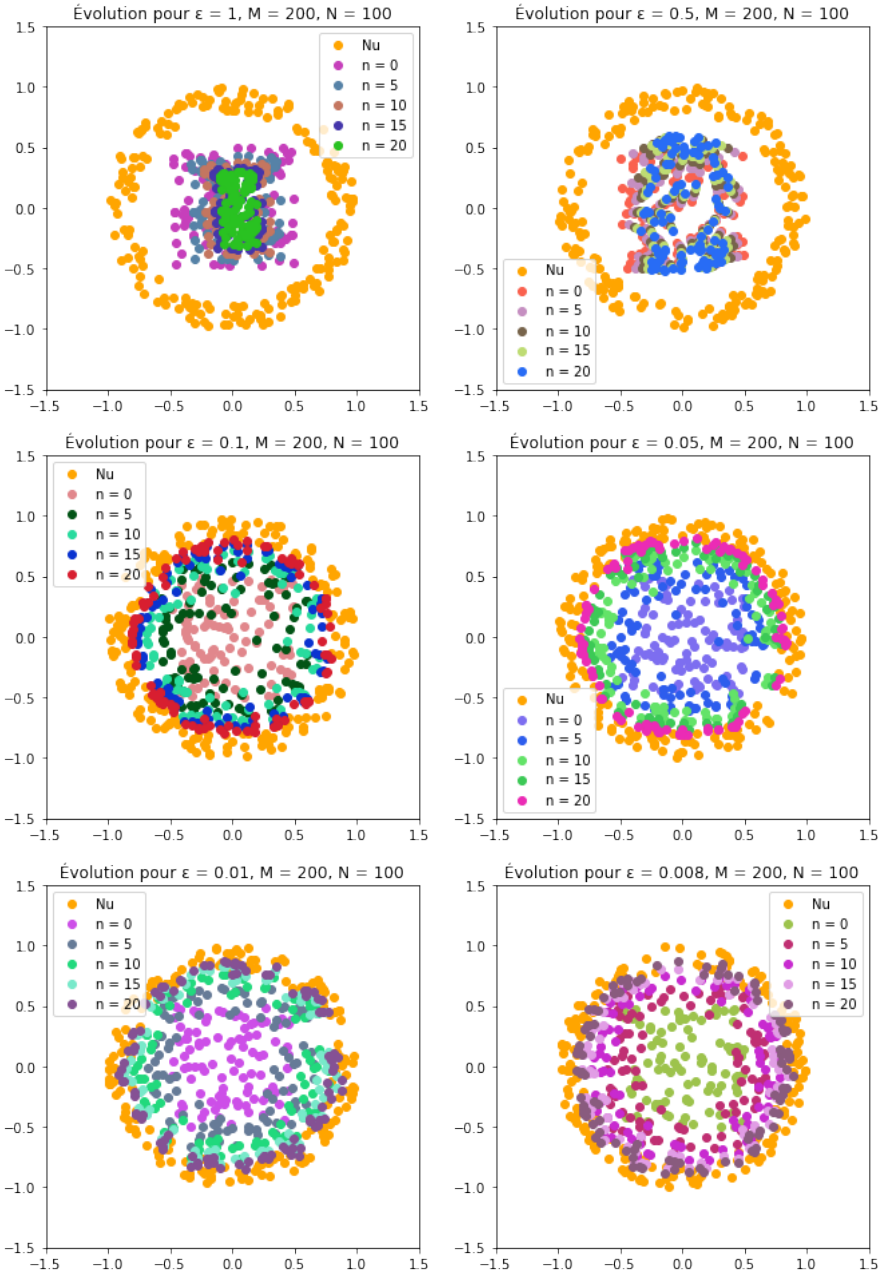


FIGURE 23

On voit une évolution identique au cas précédent. Lorsque ε est grand, la mesure μ se concentre au centre, mais lorsque $\varepsilon \to 0$, on observe une convergence vers ν. Notons que dans ce cas, ce n'est pas possible d'associer à chaque point de μ un point de ν.

6.3 Divergences de Sinkhorn

6.3.1 Introduction

Une propriété notable de la "distance" de Wasserstein régularisée est qu'elle n'est pas du tout une distance :

$$\mathcal{W}_\varepsilon(\mu, \mu) \neq 0$$

D'ailleurs, le terme entropique introduit un biais dans le transport, c'est-à-dire, pour des cas avec des distributions asymétriques, il n'y a pas une bonne convergence de μ vers ν si ε est trop grand [10]. C'est pour ça qu'en [10] la divergence de Sinkhorn S_ε est introduite : cette quantité n'est toujours pas une distance, mais une "divergence", un type de mesure de distance très fréquente en statistiques. Une divergence satisfait

$$S_\varepsilon(\mu, \mu) = 0 \implies \mu = 0$$

mais pas nécessairement l'inégalité triangulaire, et c'est pour ça qu'une divergence n'est pas une distance.

Il existe aussi une autre interprétation de la divergence de Sinkhorn. On sait que quand $\varepsilon \to 0$, alors la solution trouvée pour le transport s'approche de la solution du problème original non régularisé. Quand $\varepsilon > 0$, l'application de transport devient "floue" : dans notre cas discret, ça veut dire que plus ε augmente, plus les points de μ et ν communiquent entre eux.

Il y a une interprétation physique à ça, car Schrödinger a aussi fait une régularisation entropique pour résoudre un problème d'intéraction de particules [11]. Dans ce problème, le paramètre ε de régularisation peut être vu comme la température. Physiquement, plus la température augmente, plus les particules interagissent. Une brève discussion sur cela est disponible en [12].

Dans ce point de vue physique, il est donc naturel de se demander quel est le comportement limite à haute température. Dans le problème de transport optimal, pour analyser ce comportement asymptotique, on utilise une norme appelé MMD, qui a des bonnes propriétés car elle est euclidienne et la complexité numérique d'utilisation est faible. Cependant, elle a aussi des mauvaises propriétés, car on perd la géométrie de la distance originale. [10]

Une manière de comprendre la divergence de Sinkhorn est de la voir comme une interpolation entre ces deux comportement limites $\varepsilon \to 0, \varepsilon \to \infty$. C'est-à-dire, elle est une quantité qui permet de préserver la géométrie originale, et d'avoir une complexité algorithmique suffisamment basse pour être utilisée dans différentes applications.

6.3.2 Utilisation de la divergence de Sinkhorn dans l'exemple 6.2.2

Comme avant, on doit calculer le gradient de la divergence de Sinkhorn. Or, dans la divergence de Sinkhorn, le terme $\mathcal{W}_\varepsilon(\nu, \nu)$ ne dépend pas de μ, donc sa dérivée par rapport à la coordonnée $\mu_i^{(x,y)}$ est zéro. Le terme $\mathcal{W}_\varepsilon(\mu, \nu)$ est le terme utilisé dans la section précédente, et on a déjà calculé son gradient. On voit donc qu'il faut juste calculer le gradient de $\mathcal{W}_\varepsilon(\mu, \mu)$.

On peut utiliser la même formule que précédemment, en remarquant qu'à cause du terme entropique dans la régularisation, le transport optimal de μ vers μ n'est pas nécessairement l'identité. On calcule normalement γ avec Sinkhorn. C'est-à-dire, si l'on note par $\gamma^{\mu,\mu}$ la solution trouvée par Sinkhorn avec μ et μ comme argument,

$$\nabla_z \mathcal{W}_\varepsilon(\mu,\mu) = \gamma^{\mu,\mu} \cdot \nabla C^{\mu,\mu}(z)$$

La formule complète est donc

$$\nabla S_\varepsilon(\mu,\nu) = \gamma^{\mu,\nu} \cdot \nabla C^{\mu,\nu}(z) - \frac{1}{2}\gamma^{\mu,\mu} \cdot \nabla C^{\mu,\mu}(z)$$

où en particulier on note que $\nabla C^{\mu,\mu}$ n'est pas zéro. Par exemple,

$$\frac{\partial}{\partial \mu_i^x} C^{\mu,\mu} = \mu_i^x - \mu_j^x \neq 0 \quad \text{en général}$$

On peut implémenter ce gradient et reprendre le travail précédent pour la même configuration initiale et pour 20 itérations. On observe à l'itération initiale le champ gradient suivant pour ε petit

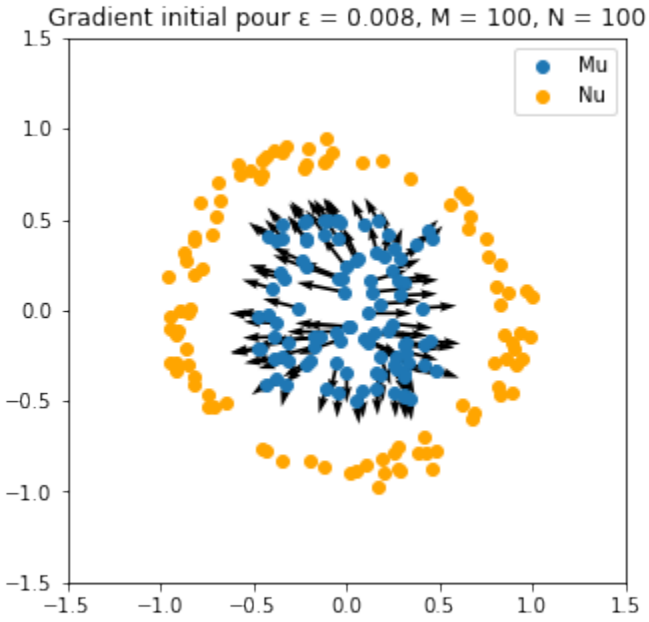


FIGURE 24

et après 20 itérations avec une descente de pas $\tau = 0.1$, lorsque ε diminue selon les mêmes valeurs que précédemment, on observe les évolutions suivantes :

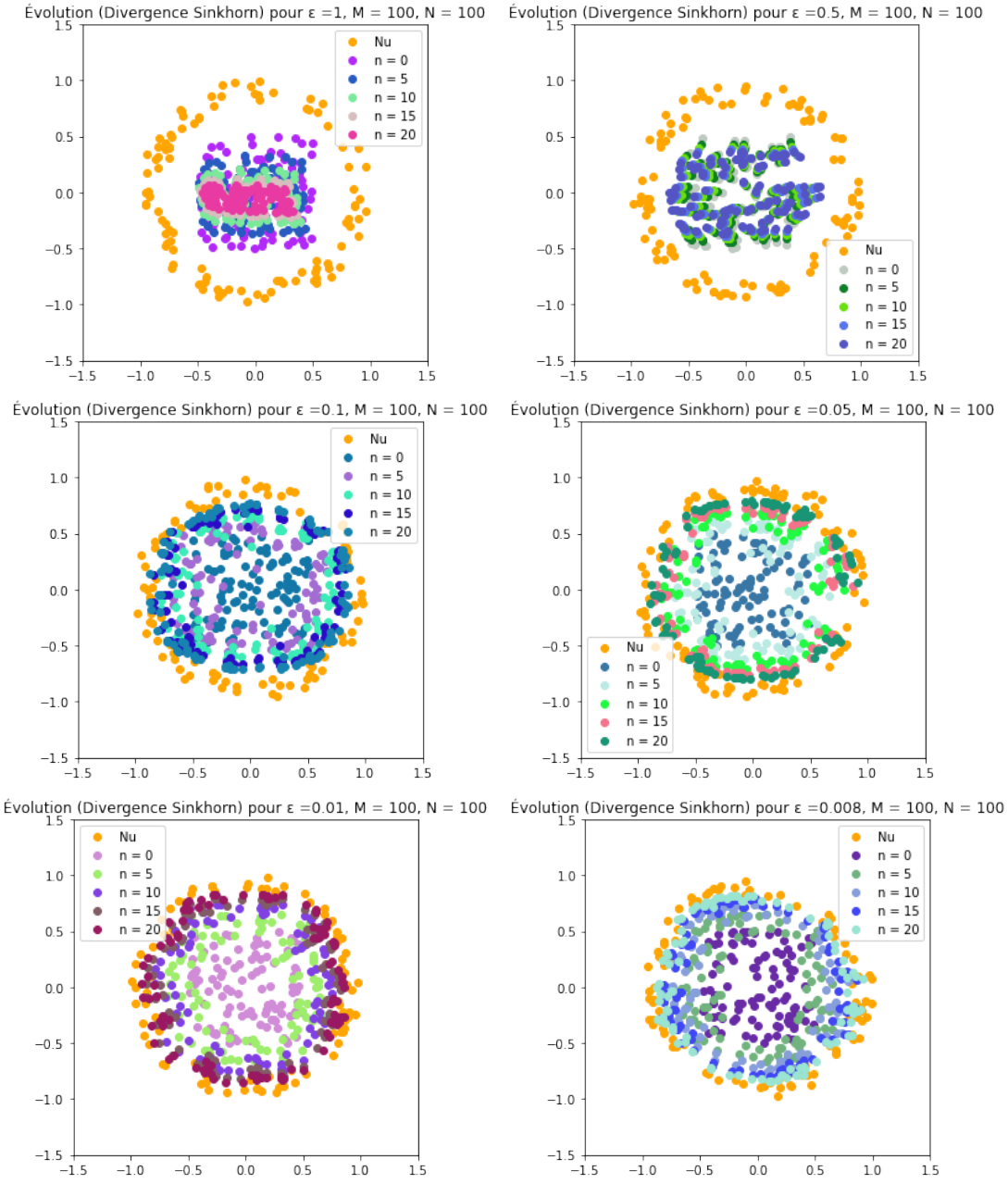


FIGURE 25

Reprenons cette expérience pour $M = 200$. On observe les évolutions suivantes :

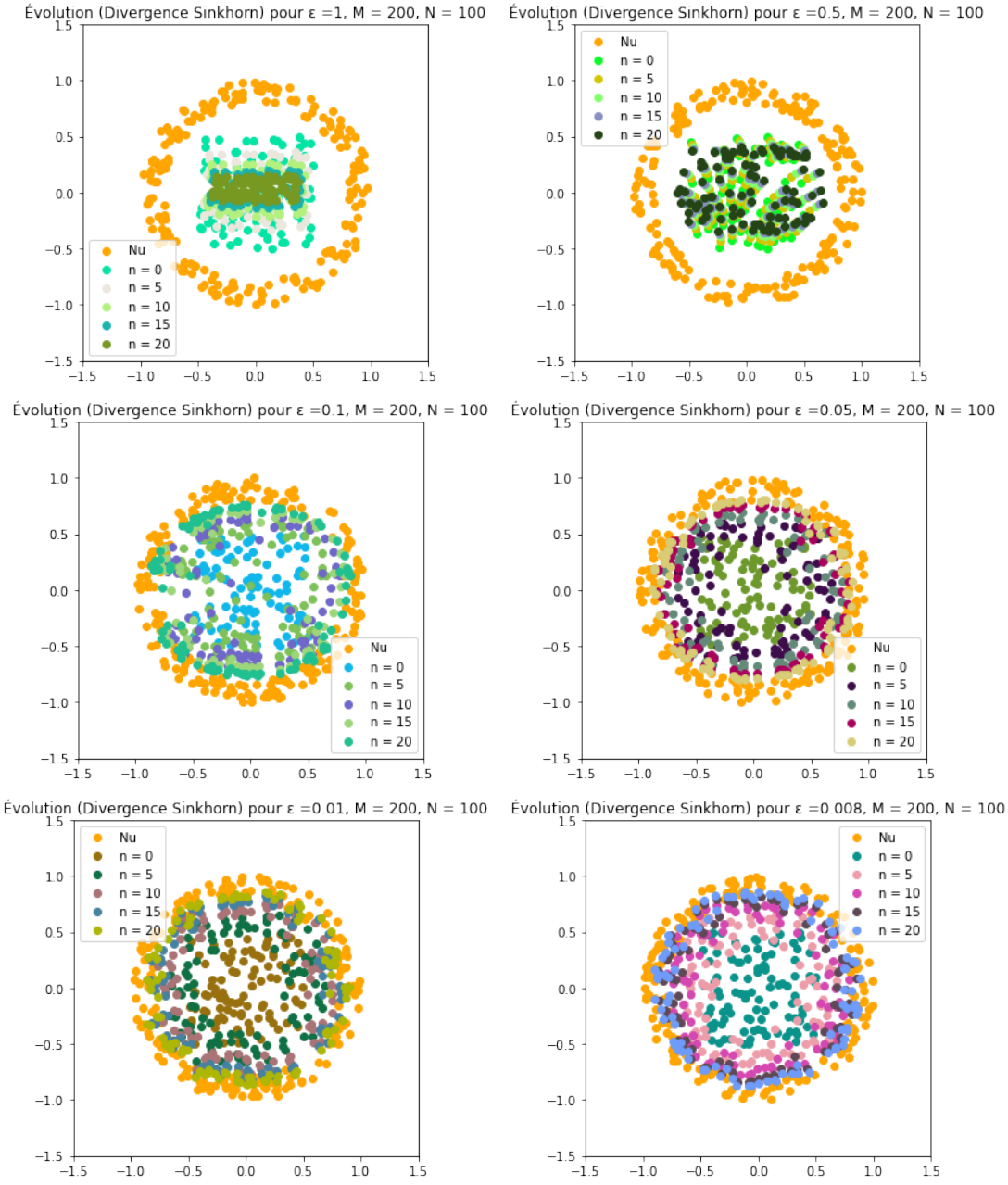


FIGURE 26

On observe une évolution similaire au cas précédent. En pratique la convergence pour ε petit se produit plus rapidement. On croit que ça arrive car doubler le nombre de points doit augmenter la norme du gradient (on peut imaginer les points ν comme des "aimants" pour les points de μ).

7 Conclusion

Dans ce rapport, nous avons exploré les principaux résultats basiques de la théorie du transport optimal, en théorie et en pratique.

D'abord, on a montré comment le problème de Monge-Kantorovich peut être formulé comme un programme linéaire et on l'a résolu en utilisant l'approche usuelle à ce type de problème, c'est-à-dire, avec l'algorithme du simplex.

En notant que l'algorithme du simplex ne marche pas bien pour des problèmes de grandes dimensions, on a étudié diverses manières de régulariser le problème. Notamment, on a utilisé la régularisation utilisant la fonction $log - sum - exp$ et la régularisation entropique. Cette dernière est devenu très populaire récemment dans la communauté du Machine Learning.

Finalement, on a adopté une approche variationnelle pour résoudre le problème de matching. On a utilisé la "distance" de Wasserstein entropiquement régularisé, et la divergence de Sinkhorn. Notons que cette dernière approche est très récente dans la littérature.

Les études réalisées dans ce projet nous ont préparé à comprendre les aspects plus récentes dans la recherche en transport optimal. C'est un sujet très vaste et qui croît rapidement aujourd'hui. On peut notamment étudier les thèmes de transport optimal "non-équilibré" [13], le transport optimal multi-marges [1], les divergences de Sinkhorn [10] et les distances de Gromov-Wasserstein [14][15].

Références

[1] Luca Nenna. "Numerical methods for multi-marginal optimal transportation". Thèse de doct. 2016.

[2] S. Kolouri et al. "Optimal Mass Transport : Signal processing and machine-learning applications". In : *IEEE Signal Processing Magazine* 34.4 (2017), p. 43-59. doi : 10.1109/MSP.2017.2695801.

[3] Marco Cuturi Gabriel Peyré. *Computational Optimal Transport*. International series of monographs on physics. Clarendon Press, 1981. isbn : 9780198520115.

[4] Filippo Santambrogio. "Optimal Transport for Applied Mathematicians. Calculus of Variations, PDEs and Modeling". In : (2015). url : https://www.math.u-psud.fr/~filippo/OTAM-cvgmt.pdf.

[5] Gaspard Monge. *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale, 1781.

[6] Étienne Ghys. *CNRS Images des Mathématiques : Gaspard Monge*. url : https://images.math.cnrs.fr/Gaspard-Monge,1094.html?lang=fr.

[7] Marco Cuturi. *Sinkhorn Distances : Lightspeed Computation of Optimal Transportation Distances*. 2013. arXiv : 1306.0895 [stat.ML].

[8] Richard Sinkhorn Paul Knopp. "Concerning nonnegative matrices and doubly stochastic matrices". In : *Pacific J. Math.* 21 (1967), p. 343-348.

[9] Giulia Luise et al. *Differential Properties of Sinkhorn Approximation for Learning with Wasserstein Distance*. 2018. arXiv : 1805.11897 [stat.ML].

[10] Jean Feydy et al. "Interpolating between Optimal Transport and MMD using Sinkhorn Divergences". working paper or preprint. Oct. 2018. url : https://hal.archives-ouvertes.fr/hal-01898858.

[11] Erwin Schrödinger. "Über die Umkehrung der Naturgesetze (in German)". In : *Preuss. Akad. Wiss. Berlin. Phys. Math.* 144 (1931), p. 144-153.

[12] Gabriel Peyré. *Optimal transport for machine learning*. url : https://www.youtube.com/watch?v=mITml5ZpqM8.

[13] Lenaïc Chizat. "Unbalanced Optimal Transport : Models, Numerical Methods, Applications". Theses. Université Paris sciences et lettres, nov. 2017. url : https://tel.archives-ouvertes.fr/tel-01881166.

[14] Titouan Vayer et al. *Sliced Gromov-Wasserstein*. 2020. arXiv : 1905.10124 [stat.ML].

[15] Titouan Vayer et al. *Fused Gromov-Wasserstein distance for structured objects : theoretical foundations and mathematical properties*. 2018. arXiv : 1811.02834 [stat.ML].