Le transport optimal numérique

Guillaume Lambert, Leonardo Martins Bianco
Enseignant : Luca Nenna

Février 2021

Introduction du problème

Gaspard Monge, *Mémoire sur la théorie des déblais et de remblais*, 1781

Léonid Kantorovich, *On the translocation of masses*, 1944

Gaspard Monge

Léonid Kantorovich

Introduction du problème

Problème de Monge-Kantorovich

Soit $C \in \mathcal{M}_{n,n}(\mathbb{R})$ une matrice coût. Le problème de Monge-Kantorovich s'écrit sous la forme

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

où

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

Résolution par le simplex

Formulation sous forme standard

On peut réécire $(\mathcal{MK})$ sous la forme standard

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

Le problème dual s'écrit donc

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

Problème :

La résolution par l'algorithme du simplex est coûteuse

Problème :
La résolution par l'algorithme du simplex est coûteuse

Solution :
Méthodes de régularisation des problèmes

Problème :
La résolution par l'algorithme du simplex est coûteuse

Solution :
Méthodes de régularisation des problèmes

Propriétés recherchées lors de la régularisation :

- approximation de la fonction-objectif de base
- le problème régularisé est sans contraintes
- convexité de la nouvelle fonction-objectif

Déroulement de l'oral

Comment résoudre le problème de Monge-Kantorovich grâce à la régularisation entropique ?

Déroulement de l'oral

Comment résoudre le problème de Monge-Kantorovich grâce à la régularisation entropique ?

Plan :

- Introduction de la régularisation entropique

- Etude du problème régularisé
  - Propriétés intéressantes
  - Equations de Bernstein-Schrödinger
  - Dualisation du problème

- Application à un exemple
  - Algorithme de gradient
  - Algorithme de Sinkhorn

La régularisation entropique

On reprend le problème primal $(\mathcal{MK})$ et on ajoute un terme d'entropie régularisé par $\varepsilon > 0$ à sa fonction-objectif. On obtient

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

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

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

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

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

Propriétés intéressantes de cette régularisation :

- F est **différentiable**.
- F est **convexe**.
- L'ensemble admissible est **convexe**.

$\Longrightarrow$ Le problème régularisé est donc **convexe**.

Résumé de la preuve que F est convexe

Point-clefs de la preuve :

Résumé de la preuve que F est convexe

Point-clefs de la preuve :

- $F\left(\gamma(1-t)+t\beta\right)=\sum_{i,j} g(\gamma_{ij}(1-t)+\beta_{ij}t)$ avec

$$g : x \mapsto C_{ij}x + \varepsilon x(\log(\frac{x}{\mu_i \nu_j}) - 1)$$

Résumé de la preuve que F est convexe

Point-clefs de la preuve :

- $F\left(\gamma(1-t)+t\beta\right)=\sum_{i,j} g(\gamma_{ij}(1-t)+\beta_{ij}t)$ avec

$$g : x \mapsto C_{ij}x + \varepsilon x(\log(\frac{x}{\mu_i \nu_j}) - 1)$$

- g est convexe comme somme de 2 fonctions convexes

Résumé de la preuve que F est convexe

Point-clefs de la preuve :

- $F\left(\gamma(1-t)+t\beta\right)=\sum_{i,j} g(\gamma_{ij}(1-t)+\beta_{ij}t)$ avec

$$g : x \mapsto C_{ij}x + \varepsilon x(\log(\frac{x}{\mu_i \nu_j}) - 1)$$

- g est convexe comme somme de 2 fonctions convexes
- On termine la preuve en utilisant la convexité de g et la linéarité d'une somme finie.

Résumé de la preuve que X_E est convexe

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.

Soient $\gamma, \beta \in X_E$ et $t \in [0, 1]$. On veut montrer que $\gamma t + \beta(1 - t) \in X_E$.

Point-clefs de la preuve :

Résumé de la preuve que X_E est convexe

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.

Soient $\gamma, \beta \in X_E$ et $t \in [0, 1]$. On veut montrer que $\gamma t + \beta(1 - t) \in X_E$.

Point-clefs de la preuve :

- linéarité d'une somme finie

Résumé de la preuve que X_E est convexe

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.

Soient $\gamma, \beta \in X_E$ et $t \in [0, 1]$. On veut montrer que $\gamma t + \beta(1 - t) \in X_E$.

Point-clefs de la preuve :

- linéarité d'une somme finie
- une combinaison linéaire d'éléments positifs est positive.

Nouvelle formulation de F

Avec la notation $\bar{\gamma}_{ij} := \mu_i \nu_j \exp\left(\frac{-C_{ij}}{\varepsilon}\right)$ et après calculs, on peut reformuler F :

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

Equations de Bernstein-Schrödinger

Equations de Bernstein-Schrödinger

La solution $\gamma^\star$ peut s'écrire comme

$$\gamma^\star_{ij} = a_i b_j \overline{\gamma}_{ij} \quad \forall i, j$$

où $a_i = \exp(u_i/\varepsilon)$, $b_j = \exp(v_j/\varepsilon)$ et u_i, v_j sont les multiplicateurs de Lagrange.
Les équations de Bernstein-Schrödinger sont

$$a_i = \frac{\mu_i}{\sum_j b_j \overline{\gamma}_{ij}} \text{ et } b_j = \frac{\nu_j}{\sum_i a_i \overline{\gamma}_{ij}}$$

Résumé de la preuve

Résumé de la preuve

– Le problème régularisé est convexe.

Résumé de la preuve

- Le problème régularisé est convexe.
- La fonction-objectif F et la fonction définissant les contraintes d'égalité c_E sont différentiables sur X_E.

Résumé de la preuve

- Le problème régularisé est convexe.
- La fonction-objectif F et la fonction définissant les contraintes d'égalité 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**.

On obtient le résultat en étudiant

$$\nabla \ell(x, \lambda) = 0$$

Dualisation du problème régularisé

Le dual, noté $(\mathcal{H}_d)$, est donné par

$$\sup_{u,v} \inf_{\gamma} \ell(\gamma, u, v) = \sup_{u,v} \ell(\gamma^*, u, v) \quad \text{avec} \quad \inf_{\gamma} \ell(\gamma, u, v) = \gamma^*$$

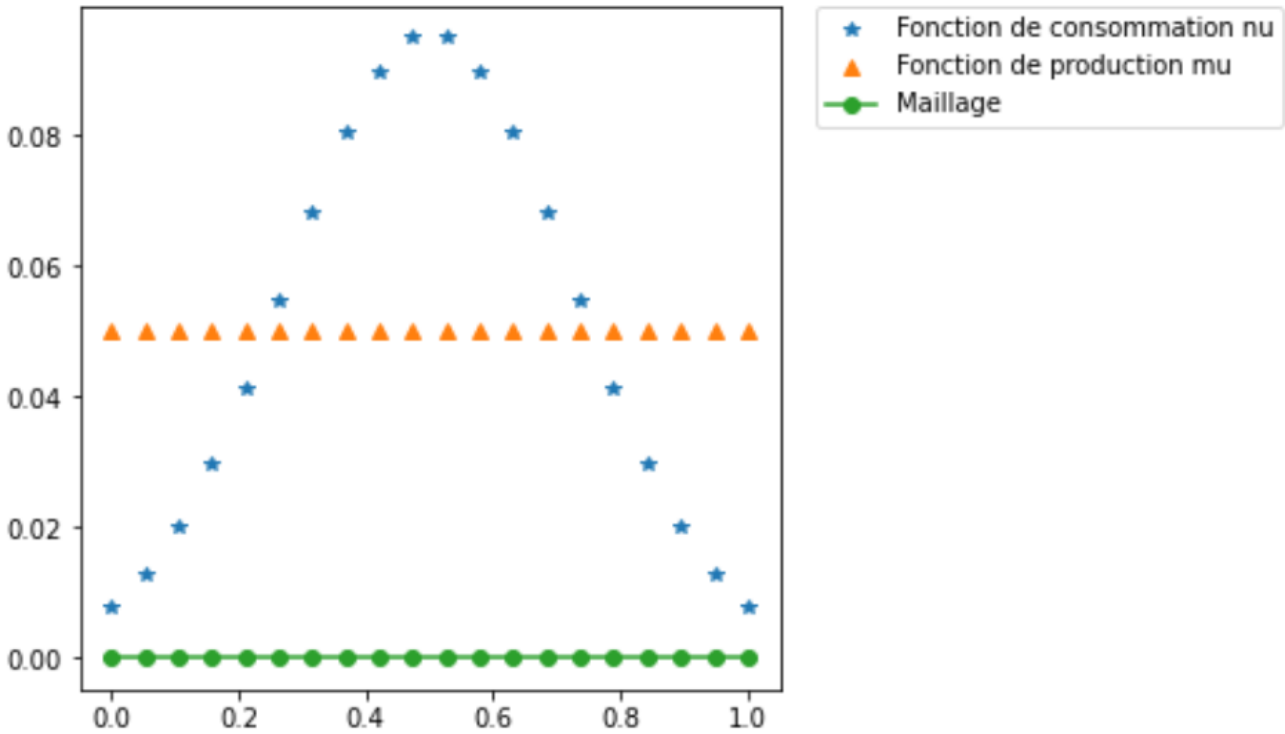
Après calculs, on a

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

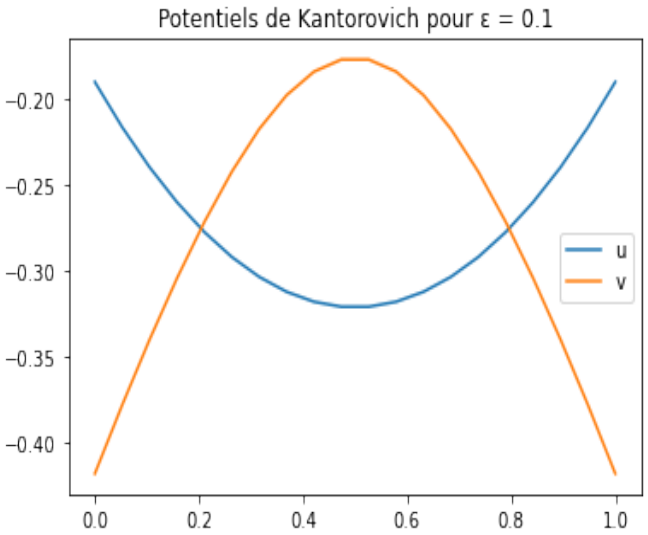
qui est un problème sans contraintes et on se ramène au problème dual $(\mathcal{MK}_d)$ en prenant $\varepsilon \to 0$

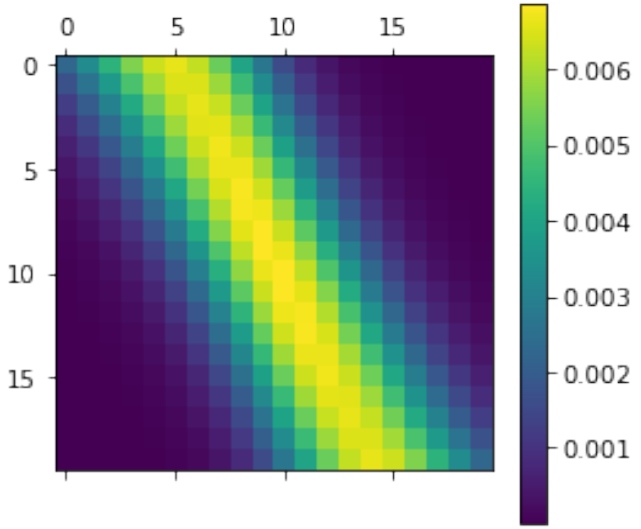
Exemple

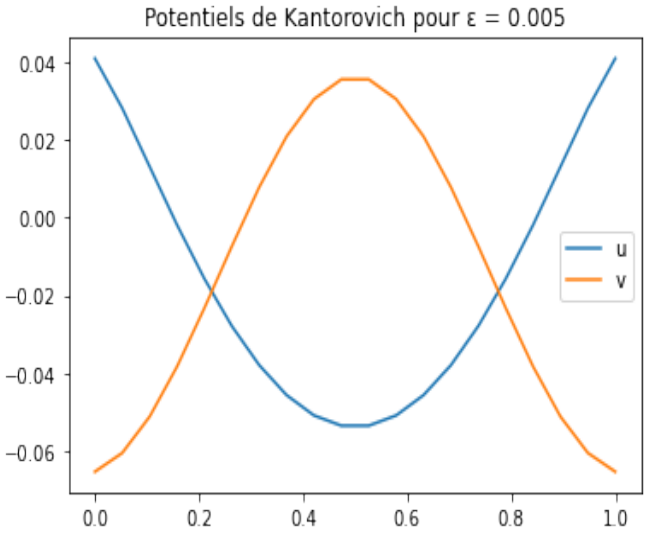
- Le domaine est l'intervalle $[0, 1]$ discrétisé par $N = 20$ points.
- La fonction-coût est la distance au carré : $c(x, y) = |x - y|^2$.
- La fonction de production est $\mu(x) = \mathbf{1}_{[0,1]}(x)$.
- La fonction de consommation est $\nu(x) = \exp(-10(y - 0.5)^2)$.

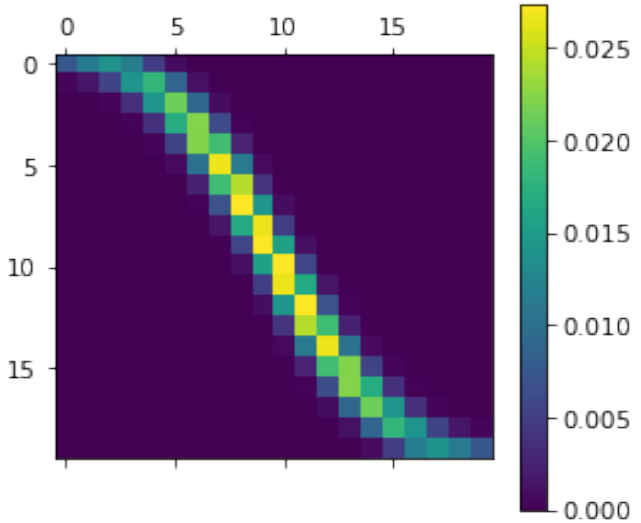


Algorithme de gradient : résultats









Algorithme de gradient : résultats

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

- l'algorithme de gradient pour (H_d) :  0.0083330246342046 7
Erreur :  0.0027792910425892443
```

Algorithme de Sinkhorn

On observe les équations de Bernstein-Schrödinger :

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

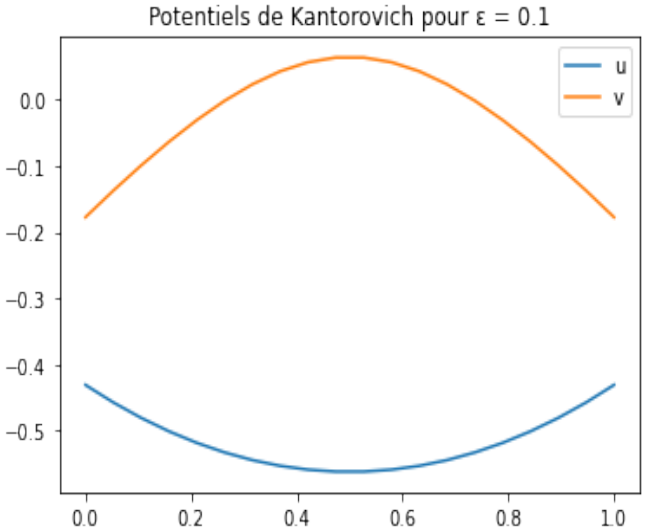
Les 2 équations sont dépendantes. Cela suggère d'écrire un algorithme itérative de la forme :

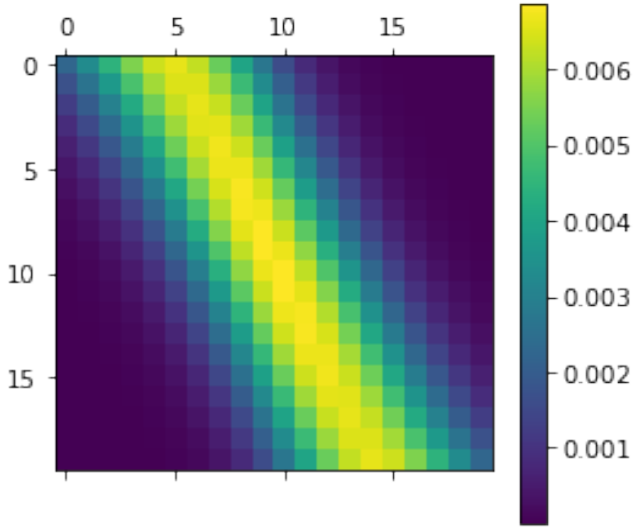
Algorithme de Sinkhorn

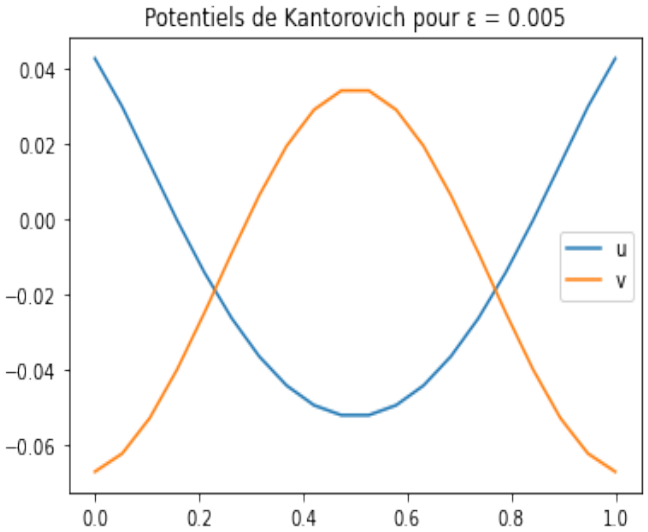
$$a_i^{n+1} = \frac{\mu_i}{\sum_j b_j^n \bar{\gamma}_{ij}} \quad b_j^{n+1} = \frac{\nu_j}{\sum_i a_i^{n+1} \bar{\gamma}_{ij}}$$

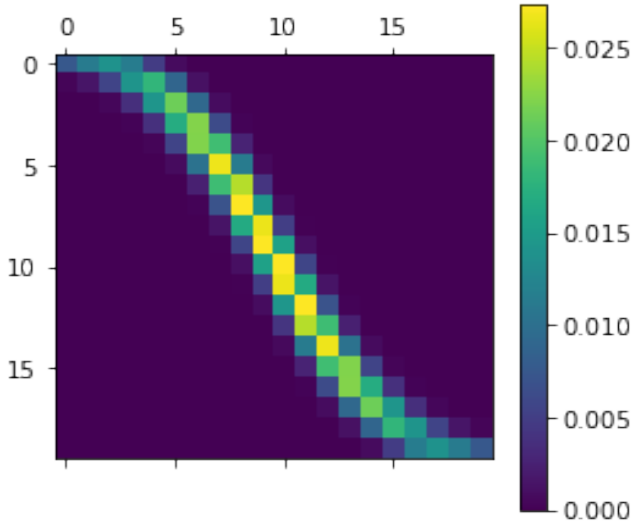
L'algorithme est appelé Sinkhorn car Richard Sinkhorn et Paul Knopp ont prouvé sa convergence.

Algorithme de Sinkhorn : résultats









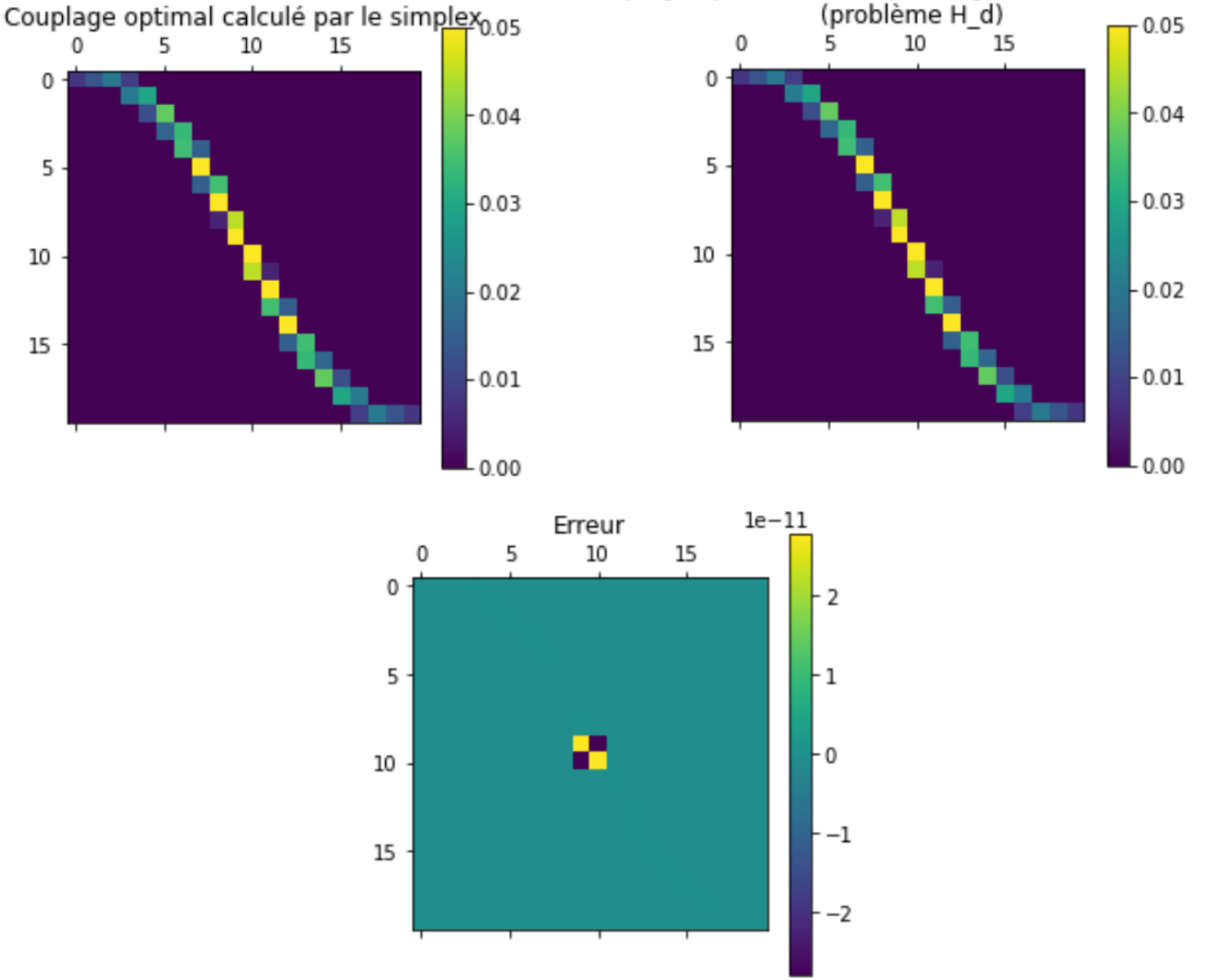
Algorithme de Sinkhorn : résultats

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

- l'algorithme de gradient pour (H_d) :  0.008333024634204668
Erreur :  0.0027792910425892443

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

Algorithme de Sinkhorn : résultats



Conclusion

En résumé :

- Le problème de Monge-Kantorovich peut être résolu par l'algorithme du simplex mais la procédure est coûteuse.
- La régularisation entropique est une méthode de régularisation permettant de palier ce problème.
- On a utilisé deux algorithmes : l'algorithme de gradient à pas fixe et l'algorithme de Sinkhorn.
- L'algorithme de Sinkhorn est plus stable et plus précis que l'algorithme de gradient.

Merci pour votre attention !

Une approche variationnelle

– La descente de gradient pour les multiplicateurs de Lagrange u et v était réalisée dans $\mathbb{R}^{2n}$.

Une approche variationnelle

- La descente de gradient pour les multiplicateurs de Lagrange u et v était réalisée dans $\mathbb{R}^{2n}$.
- On passe maintenant à une approche plus *géométrique* : on réalise cette démarche directement dans l'espace de Wasserstein (qui est non-euclidien).

Une approche variationnelle

- La descente de gradient pour les multiplicateurs de Lagrange u et v était réalisée dans $\mathbb{R}^{2n}$.
- On passe maintenant à une approche plus *géométrique* : on réalise cette démarche directement dans l'espace de Wasserstein (qui est non-euclidien).
- On va toujours chercher minimiser la distance de Wasserstein en fixant ν et variant μ. Les coordonnées de μ sont dans $\mathbb{R}^n$, mais la métrique sur cet espace est compliquée. On peut aussi utiliser la distance régularisée $\mathcal{W}_\varepsilon = \sum_{ij} C_{ij}\gamma_{ij} + \varepsilon \sum_{ij} \gamma_{ij} \left(\log\left(\frac{\gamma_{ij}}{\mu_i \nu_j}\right) - 1 \right)$.

Mise en œuvre

– Cette approche est appelée variationnelle car si l'on voit $\mathcal{W}(\mu, \nu) := \mathcal{E}(\mu)$ comme une fonctionnelle d'énergie, alors minimiser pour μ correspond à trouver géodésiques dans $(\mathbb{R}^n, \mathcal{W}_2)$.

Mise en œuvre

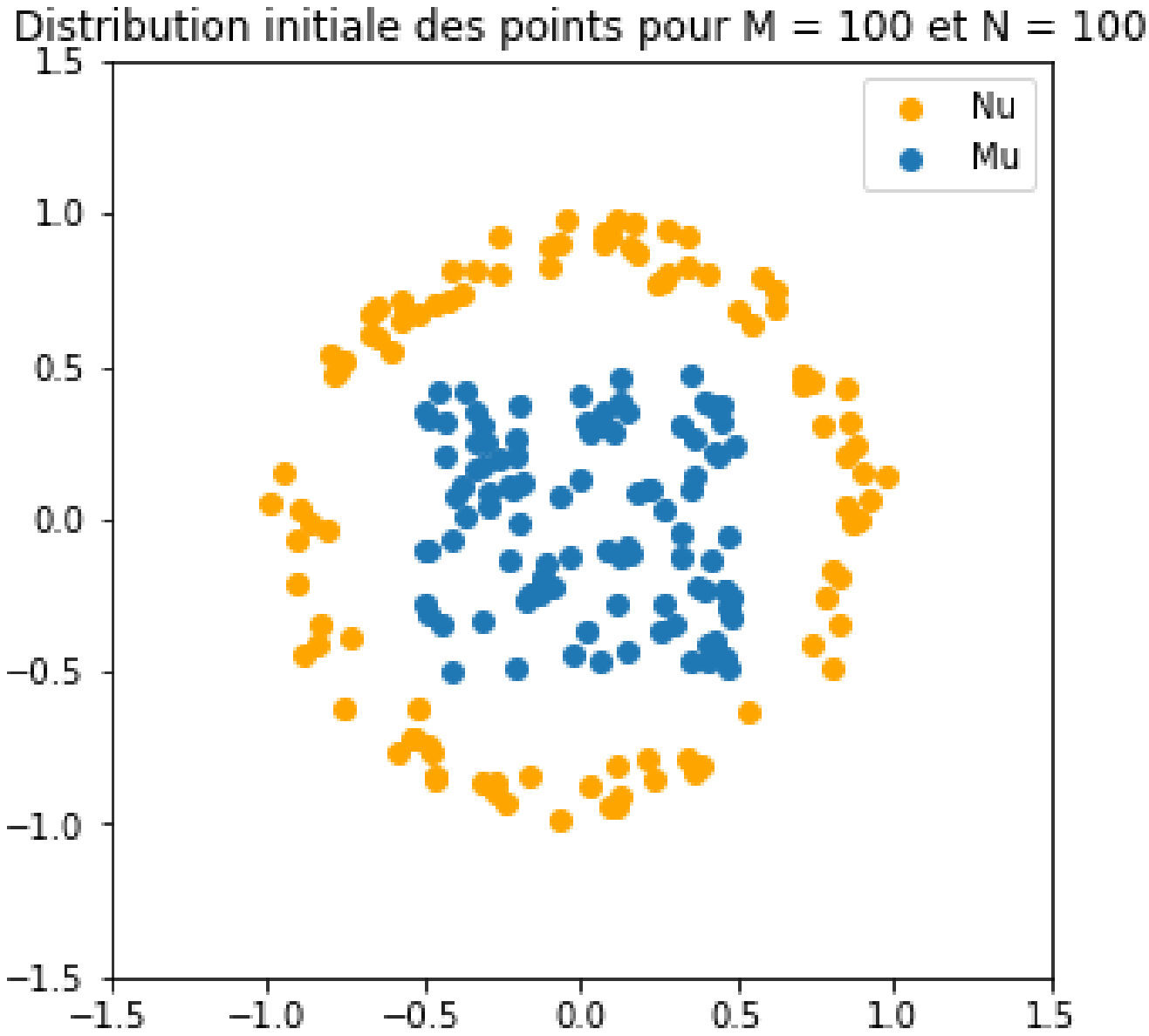
- Cette approche est appelée variationnelle car si l'on voit $\mathcal{W}(\mu, \nu) := \mathcal{E}(\mu)$ comme une fonctionnelle d'énergie, alors minimiser pour μ correspond à trouver géodésiques dans $(\mathbb{R}^n, \mathcal{W}_2)$.
- L'idée pour calculer le gradient de l'énergie est très simples : on l'écrit comme $\mathcal{E}(x) = L\left(\varepsilon, C_{ij}(x), \mu, \nu\right)$. On peut alors prendre les dérivées en utilisant la règle de dérivation en chaîne.

Mise en œuvre

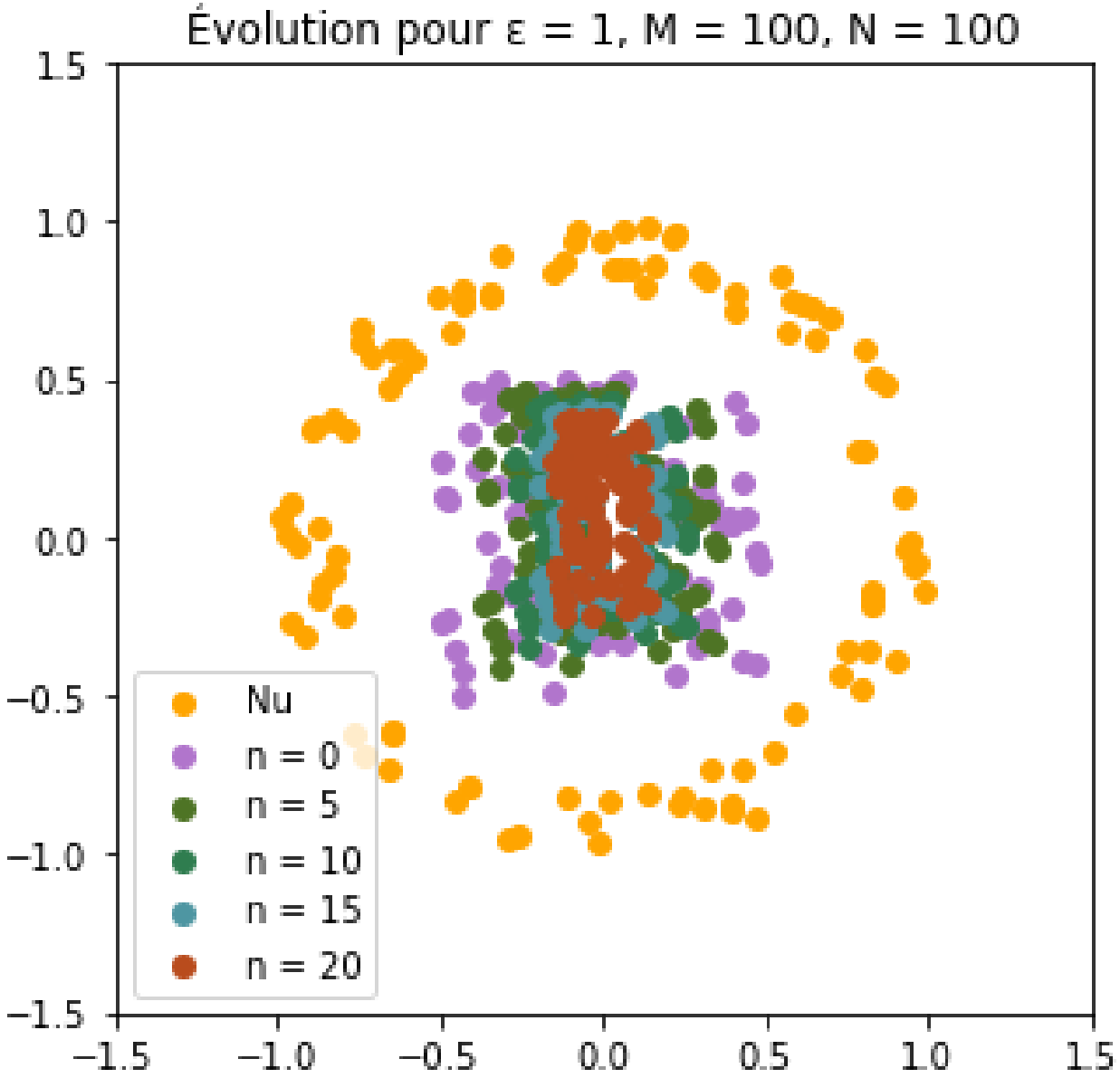
- Cette approche est appelée variationnelle car si l'on voit $\mathcal{W}(\mu,\nu) := \mathcal{E}(\mu)$ comme une fonctionnelle d'énergie, alors minimiser pour μ correspond à trouver géodésiques dans $(\mathbb{R}^n, \mathcal{W}_2)$.
- L'idée pour calculer le gradient de l'énergie est très simples : on l'écrit comme $\mathcal{E}(x) = L\left(\varepsilon, C_{ij}(x), \mu, \nu\right)$. On peut alors prendre les dérivées en utilisant la règle de dérivation en chaîne.
- On trouve $\nabla\mathcal{E}(x) = \gamma \cdot \nabla C(x)$ et en substituant le coût quadratique on obtient

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

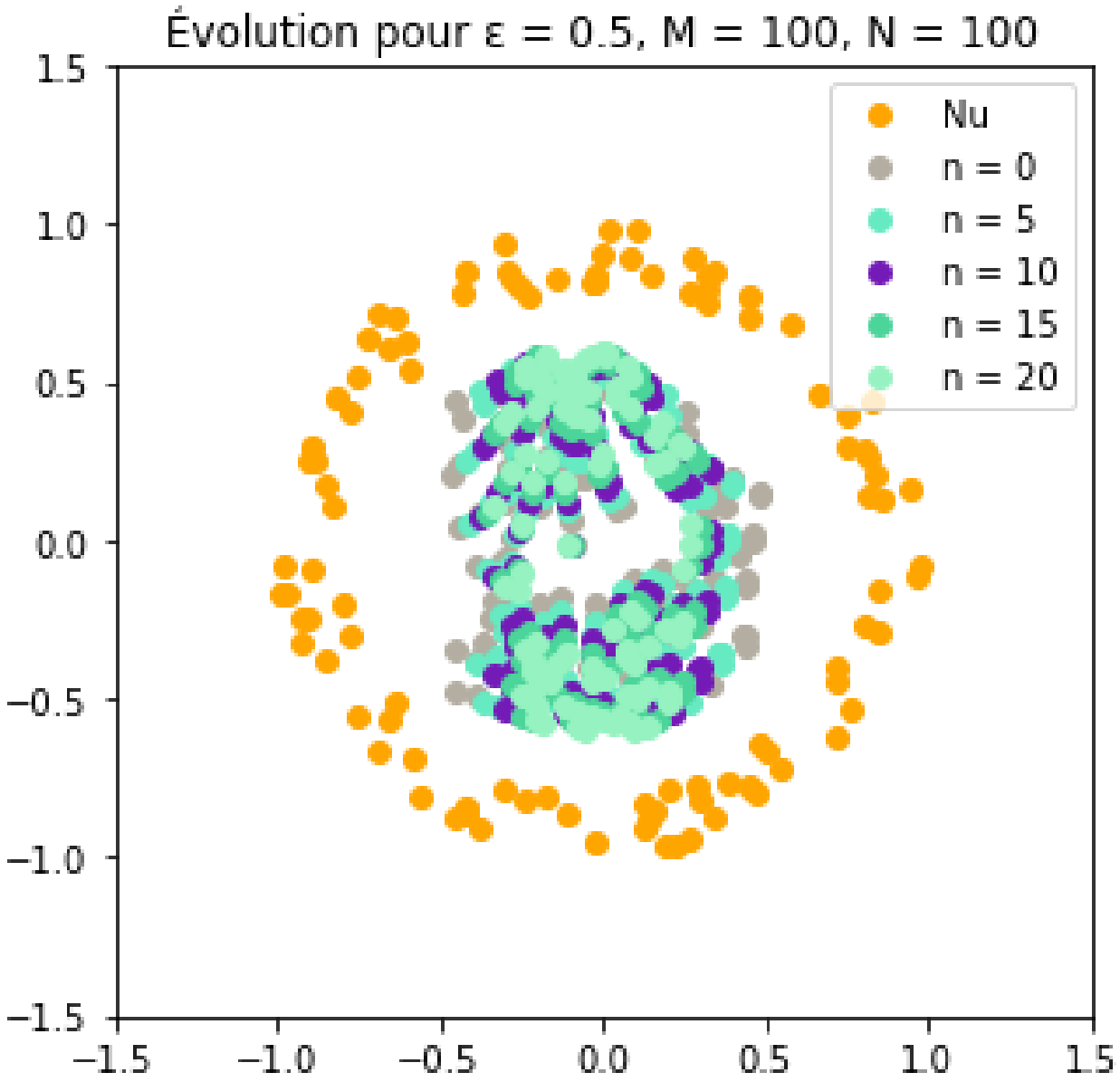
Résultats


Distribution initiale des points pour M = 100 et N = 100
Nu
Mu
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

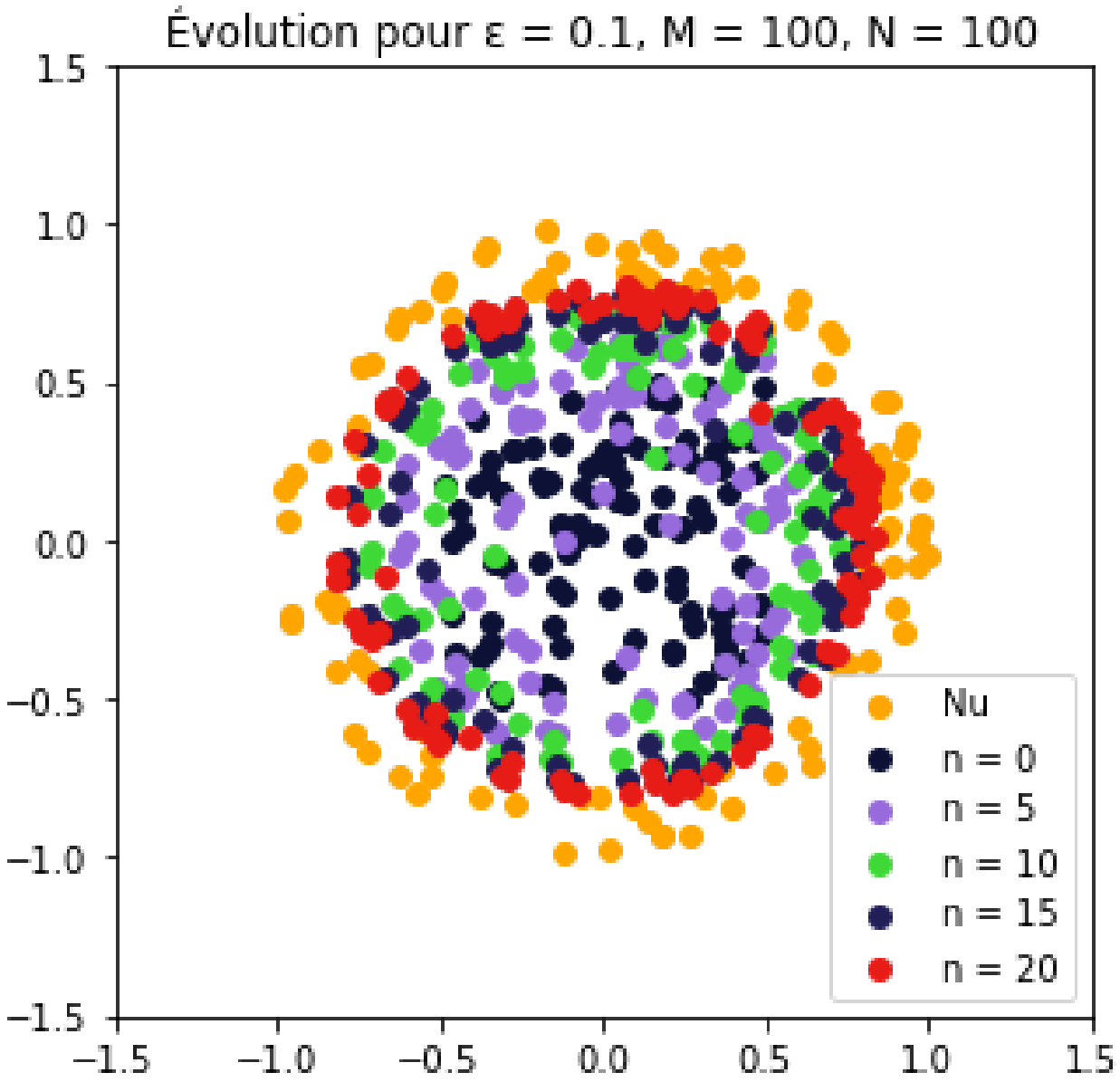
Résultats


Évolution pour ε = 1, M = 100, N = 100
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Nu
n = 0
n = 5
n = 10
n = 15
n = 20

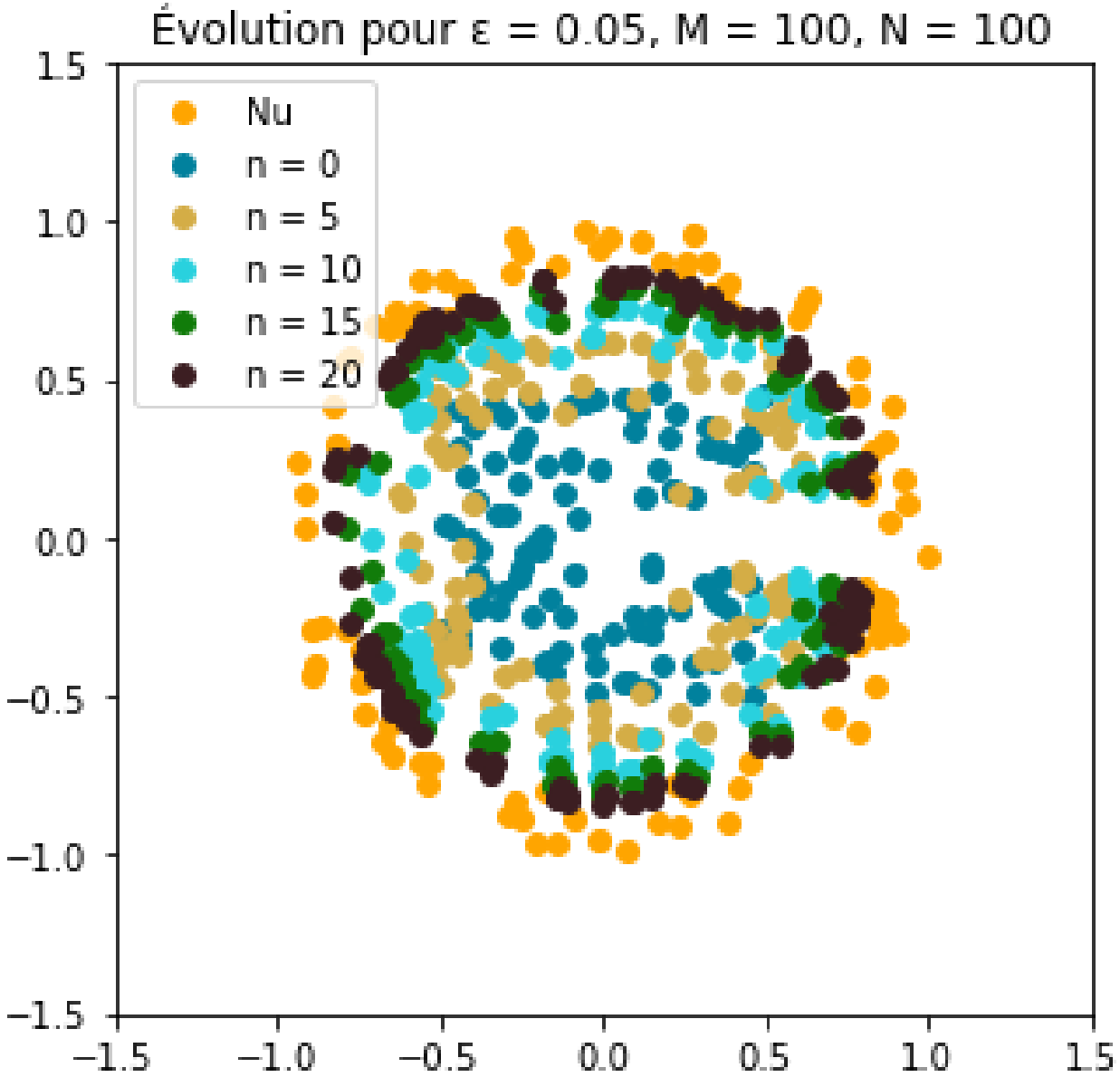
Résultats


Évolution pour ε = 0.5, M = 100, N = 100
Nu
n = 0
n = 5
n = 10
n = 15
n = 20
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

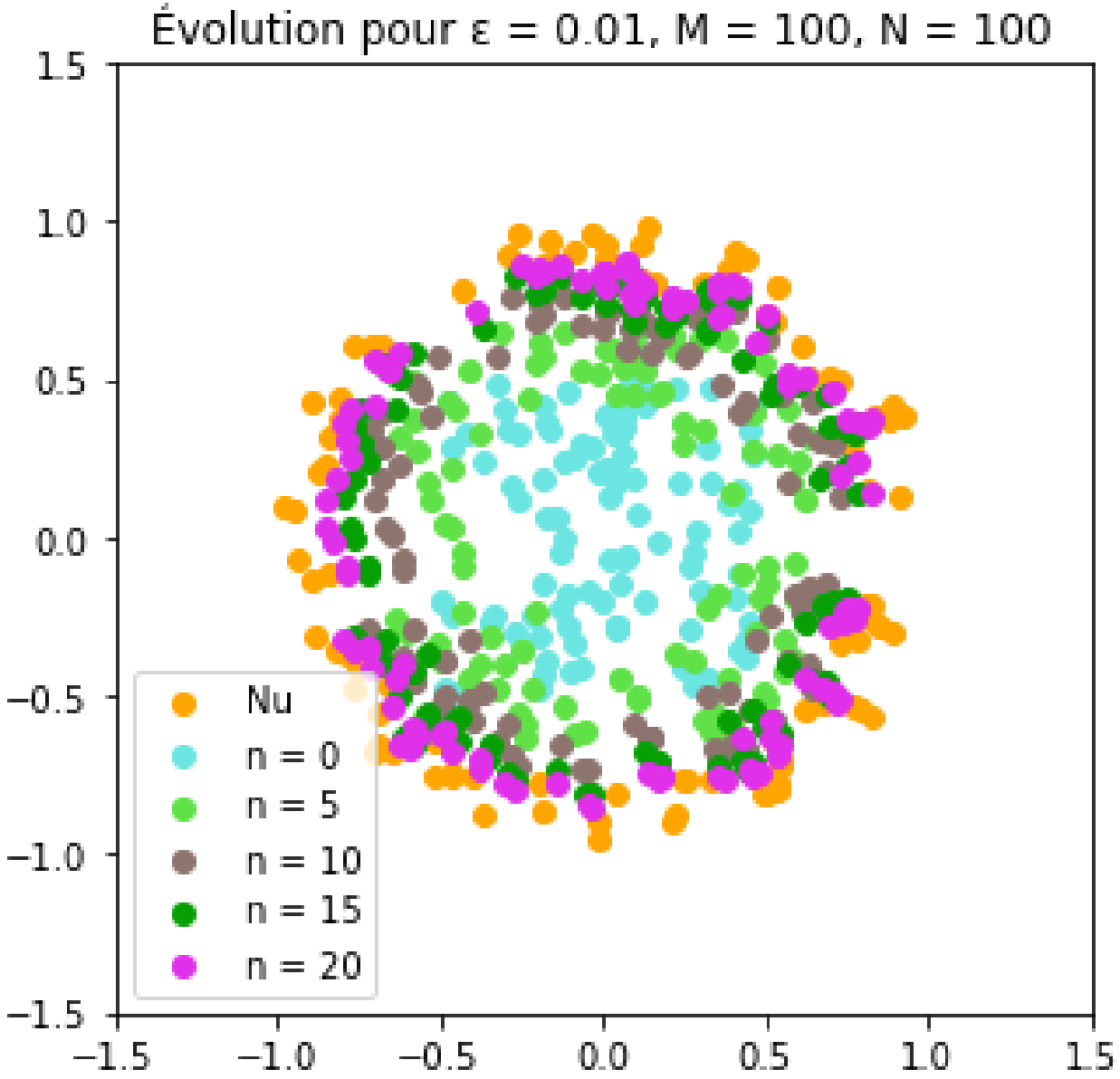
Résultats


Évolution pour ε = 0.1, M = 100, N = 100
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Nu
n = 0
n = 5
n = 10
n = 15
n = 20

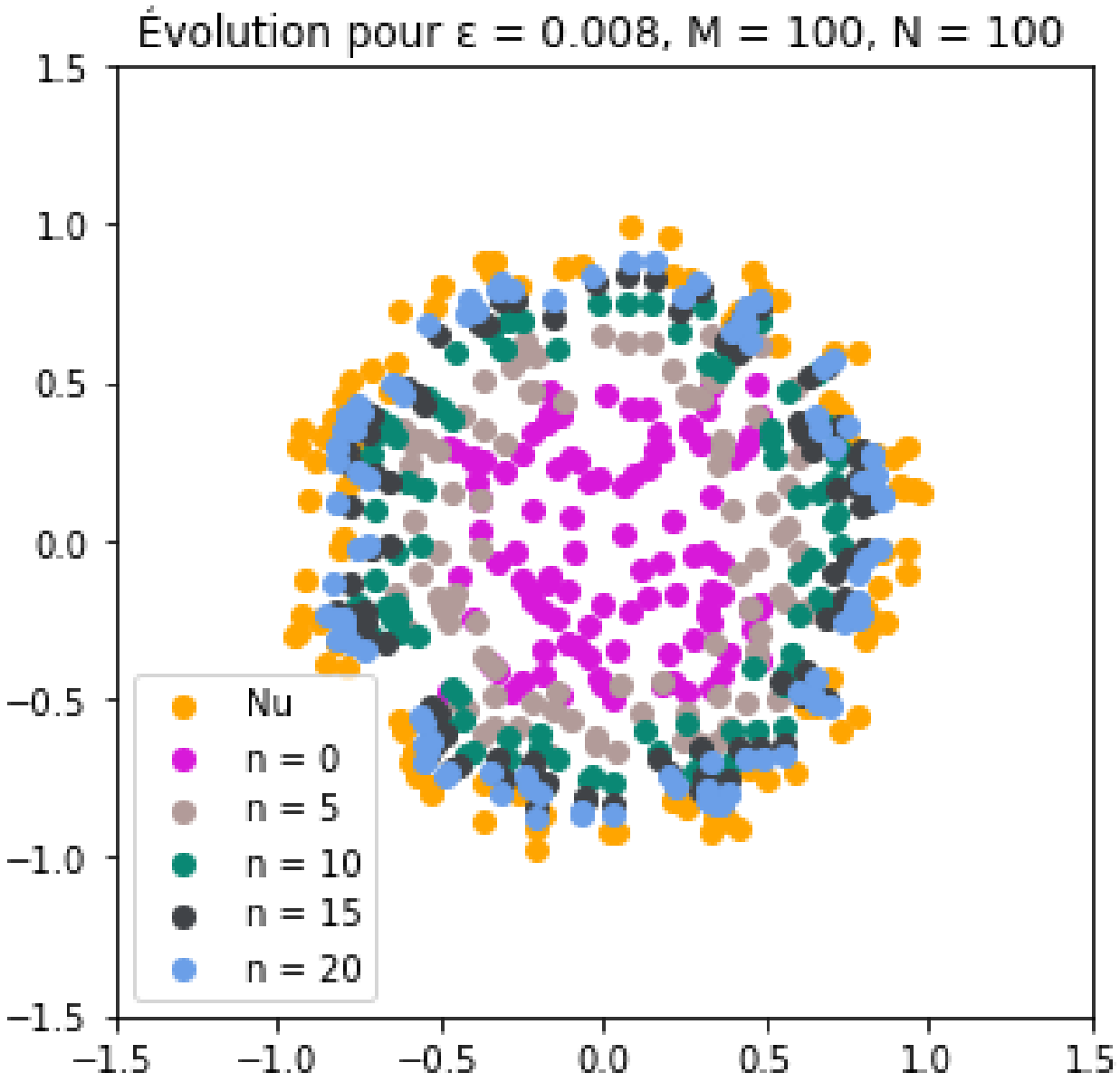
Résultats


Évolution pour ε = 0.05, M = 100, N = 100
Nu
n = 0
n = 5
n = 10
n = 15
n = 20
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Résultats


Évolution pour ε = 0.01, M = 100, N = 100
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Nu
n = 0
n = 5
n = 10
n = 15
n = 20

Résultats

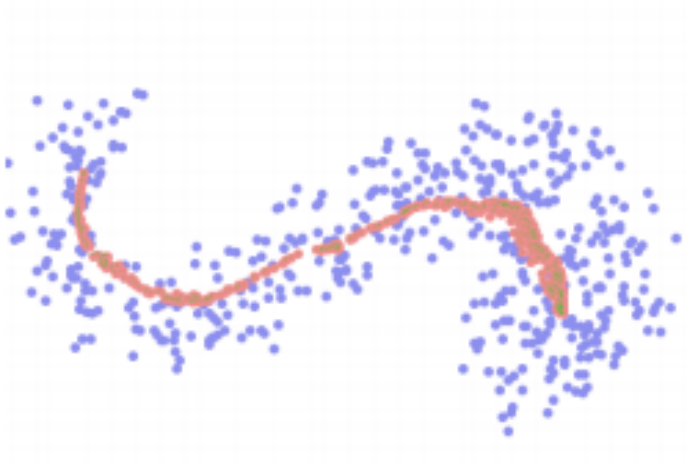

Évolution pour ε = 0.008, M = 100, N = 100
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Nu
n = 0
n = 5
n = 10
n = 15
n = 20

Problèmes

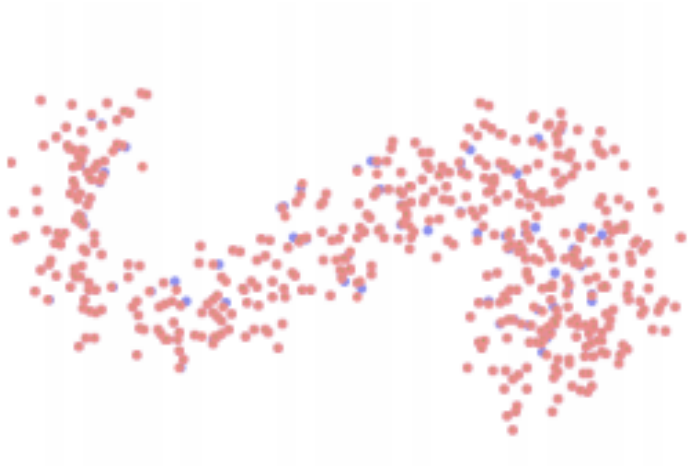
– Il arrive que le terme entropique ajoute un biais dans le transport trouvé.

Problèmes

– Il arrive que le terme entropique ajoute un biais dans le transport trouvé.



(a) $\mathrm{L} = \mathrm{OT}_\varepsilon$



(b) $\mathrm{L} = \mathrm{S}_\varepsilon$

Problèmes

– En fait $\mathcal{W}_\varepsilon$ n'est pas une distance : $\mathcal{W}_\varepsilon(\mu, \mu) \neq 0$. Cela peut être vu comme une asymétrie.

Problèmes

- En fait $\mathcal{W}_\varepsilon$ n'est pas une distance : $\mathcal{W}_\varepsilon(\mu, \mu) \neq 0$. Cela peut être vu comme une asymétrie.
- La solution proposée dans l'article récent (références dans rapport) est une "symétrisation" de la distance : on remplace $\mathcal{W}_\varepsilon$ par

$$S_\varepsilon(\mu, \nu) = \mathcal{W}_\varepsilon(\mu, \nu) - \frac{1}{2}\mathcal{W}_\varepsilon(\mu, \mu) - \frac{1}{2}\mathcal{W}_\varepsilon(\nu, \nu)$$

Problèmes

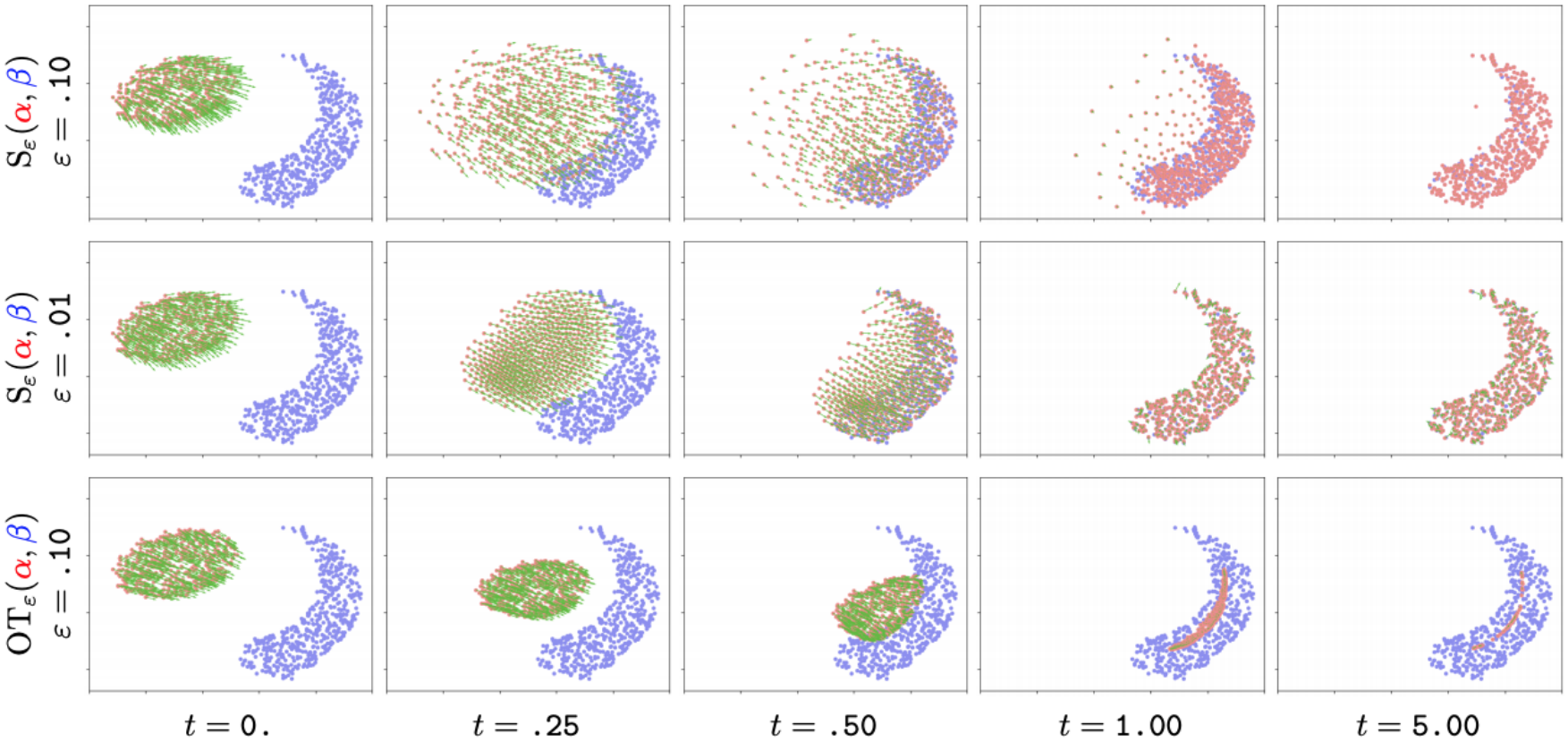
- En fait $\mathcal{W}_\varepsilon$ n'est pas une distance : $\mathcal{W}_\varepsilon(\mu, \mu) \neq 0$. Cela peut être vu comme une asymétrie.
- La solution proposée dans l'article récent (références dans rapport) est une "symétrisation" de la distance : on remplace $\mathcal{W}_\varepsilon$ par

$$S_\varepsilon(\mu, \nu) = \mathcal{W}_\varepsilon(\mu, \nu) - \frac{1}{2}\mathcal{W}_\varepsilon(\mu, \mu) - \frac{1}{2}\mathcal{W}_\varepsilon(\nu, \nu)$$

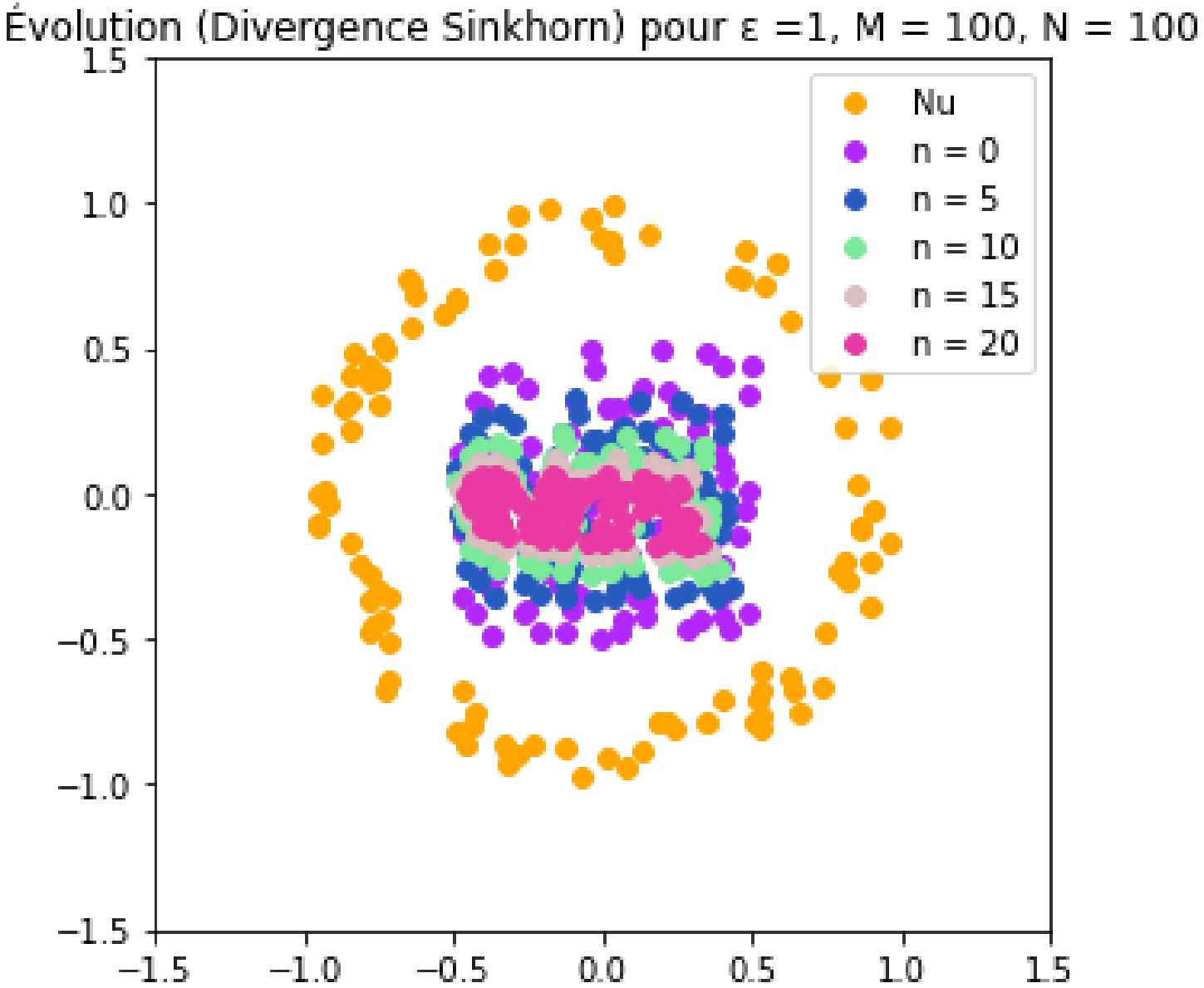
- On calcule le gradient exactement comme avant

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

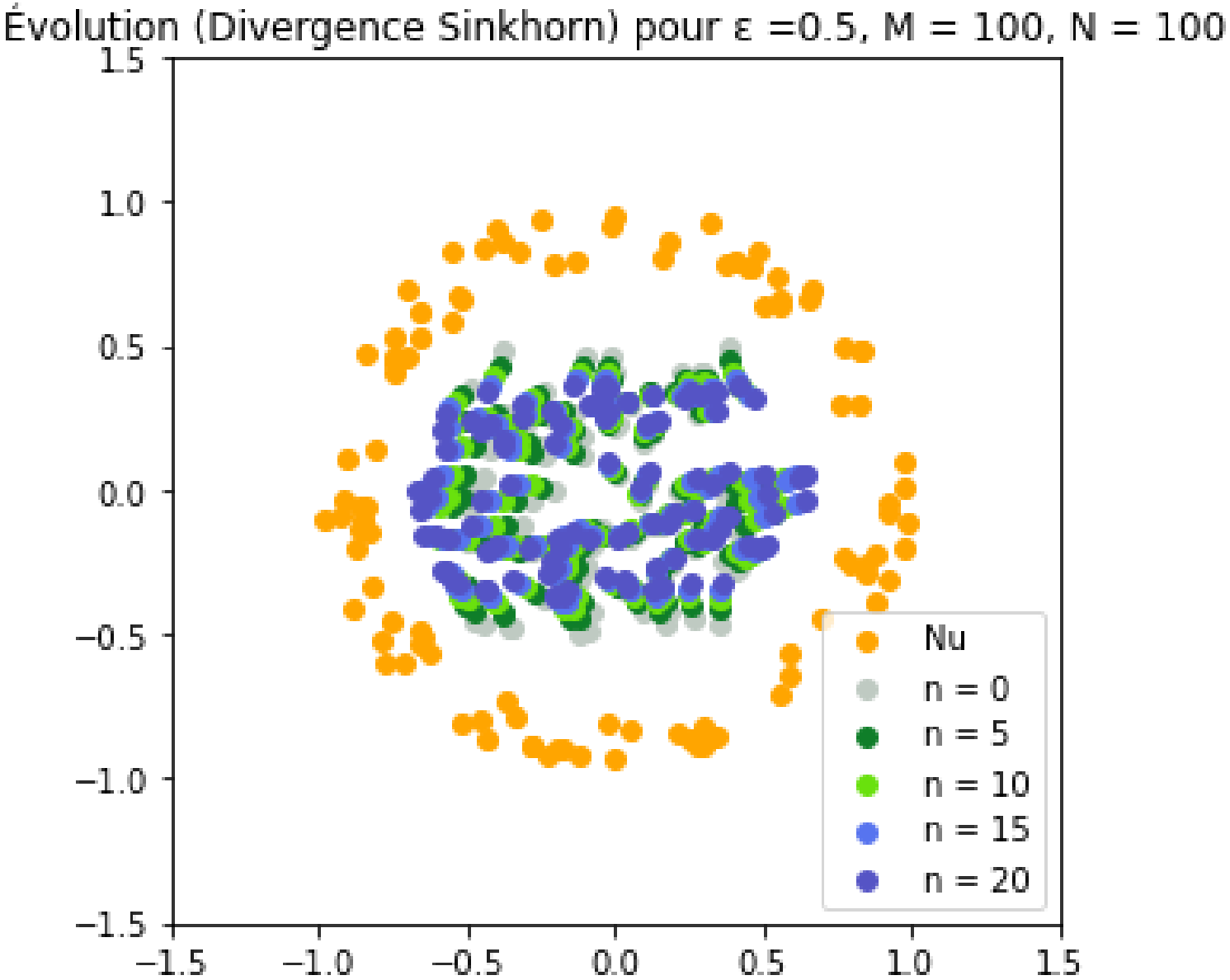
Résultats


$\mathrm{S}_\varepsilon(\alpha,\beta)$ $\varepsilon = .10$
$\mathrm{S}_\varepsilon(\alpha,\beta)$ $\varepsilon = .01$
$\mathrm{OT}_\varepsilon(\alpha,\beta)$ $\varepsilon = .10$
$t = 0.$
$t = .25$
$t = .50$
$t = 1.00$
$t = 5.00$

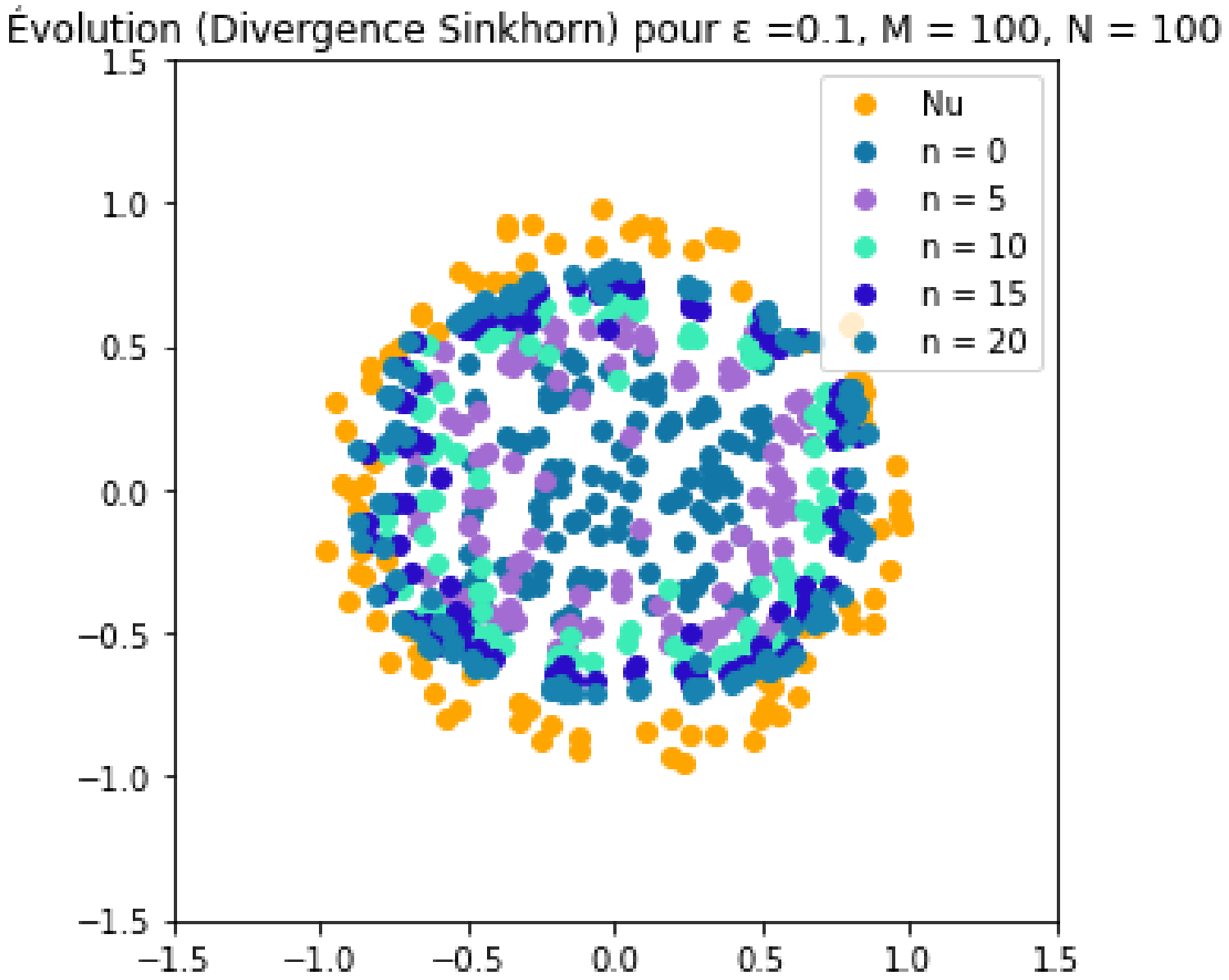
Résultats


Évolution (Divergence Sinkhorn) pour ε =1, M = 100, N = 100
Nu
n = 0
n = 5
n = 10
n = 15
n = 20
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

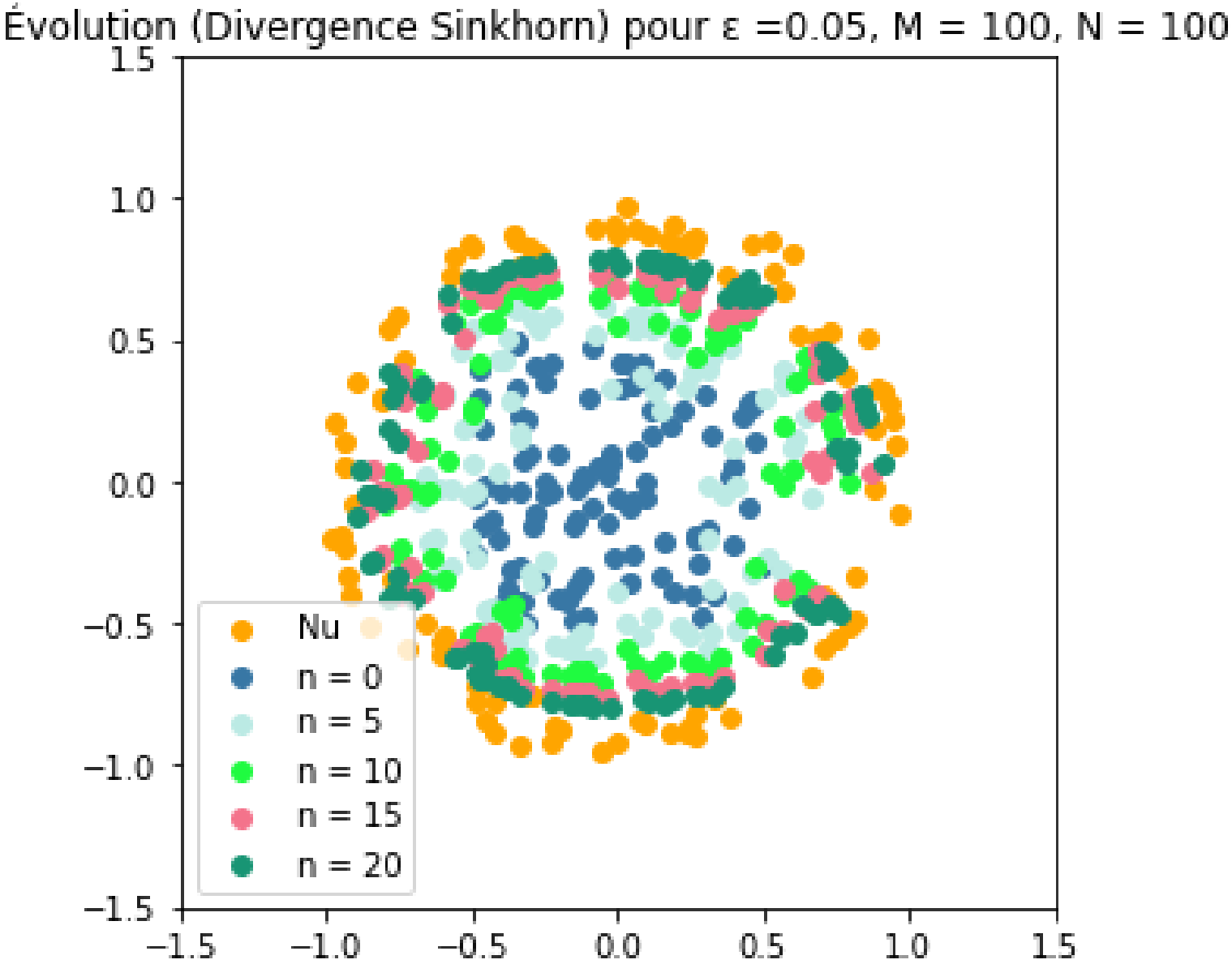
Résultats


Évolution (Divergence Sinkhorn) pour ε =0.5, M = 100, N = 100
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Nu
n = 0
n = 5
n = 10
n = 15
n = 20

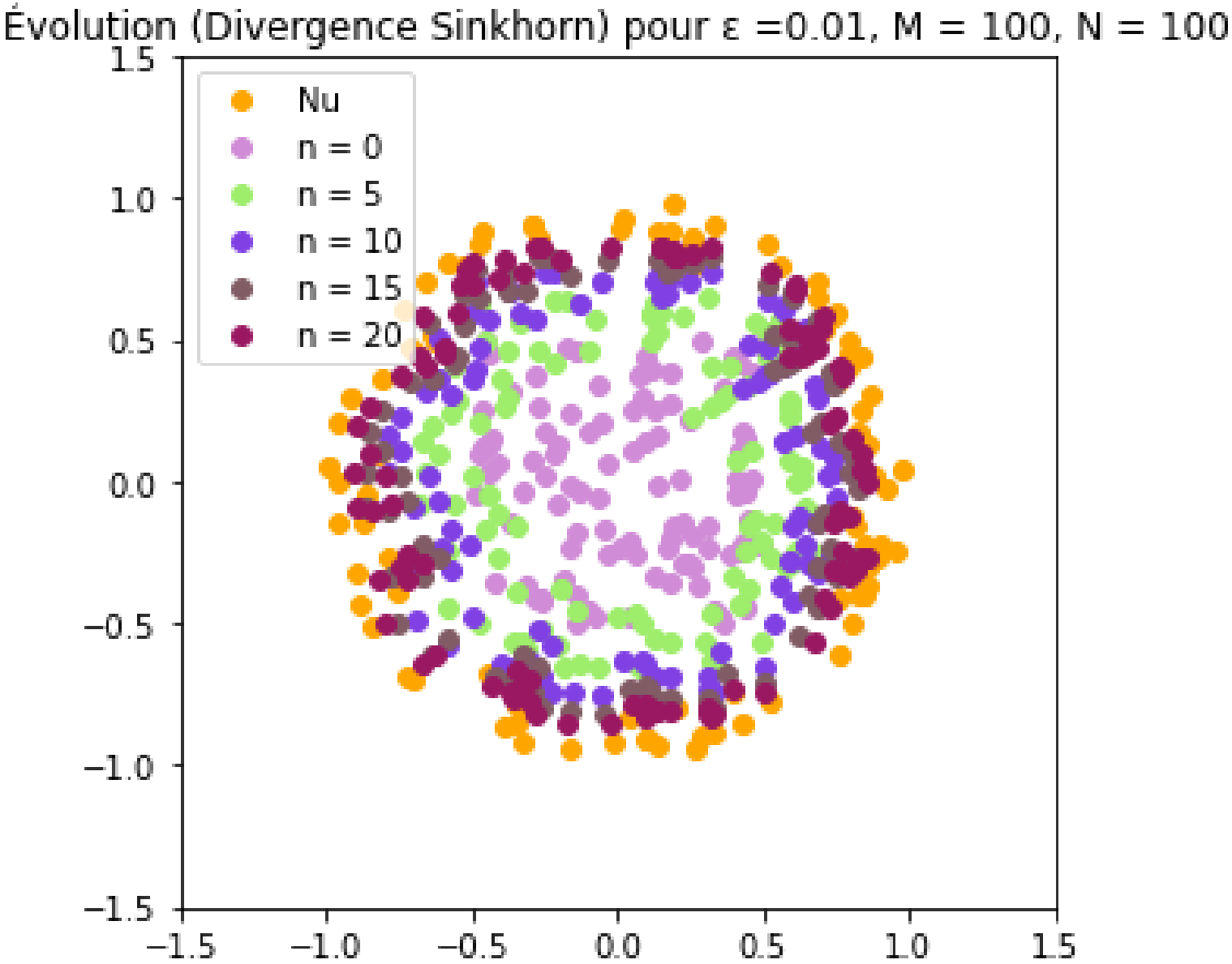
Résultats


Évolution (Divergence Sinkhorn) pour ε =0.1, M = 100, N = 100
Nu
n = 0
n = 5
n = 10
n = 15
n = 20
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

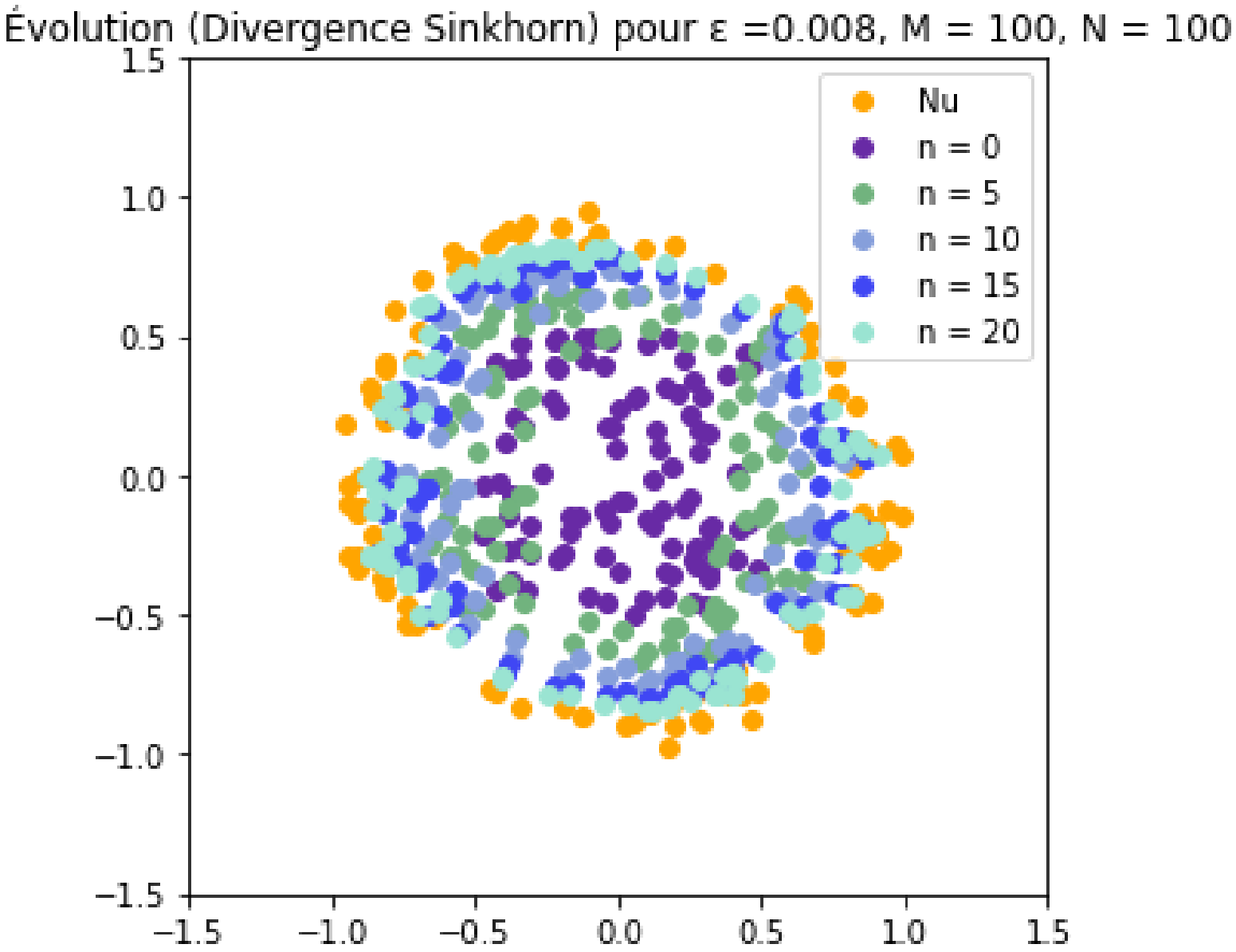
Résultats


Évolution (Divergence Sinkhorn) pour ε =0.05, M = 100, N = 100
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Nu
n = 0
n = 5
n = 10
n = 15
n = 20

Résultats


Évolution (Divergence Sinkhorn) pour ε =0.01, M = 100, N = 100
Nu
n = 0
n = 5
n = 10
n = 15
n = 20
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Résultats


Évolution (Divergence Sinkhorn) pour ε =0.008, M = 100, N = 100
Nu
n = 0
n = 5
n = 10
n = 15
n = 20
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Conclusion

– L'idée de régularisation est valide mais doit être modifiée pour devenir une divergence (bonnes propriétés).

Conclusion

- L'idée de régularisation est valide mais doit être modifiée pour devenir une divergence (bonnes propriétés).
- L'approche des divergences de Sinkhorn est de symétriser $\mathcal{W}_\varepsilon$ en utilisant une formule similaire à l'identité de polarisation.

Conclusion

- L'idée de régularisation est valide mais doit être modifiée pour devenir une divergence (bonnes propriétés).
- L'approche des divergences de Sinkhorn est de symétriser $\mathcal{W}_\varepsilon$ en utilisant une formule similaire à l'identité de polarisation.
- Recherche très récente et active!

Merci pour votre attention !