Du transport optimal et de son application en apprentissage statistique

Camille Mondon
DMA, ENS

Julien Malka
DI, ENS

14 juin 2020

Table des matières

Introduction **3**

Notations **5**

1 Le problème du transport optimal **6**
1.1 Le problème selon Kantorovitch 6
1.1.1 Formulation générale 6
1.1.2 Formulation discrète 7
1.1.3 Existence des solutions 8
1.2 Dualité lagrangienne 9

2 Les algorithmes. Régularisation entropique **11**
2.1 L'algorithme du simplexe 11
2.1.1 Sélectionner un point de départ pour notre algorithme . . . 14
2.1.2 Vérifier qu'une solution est ou non optimale 15
2.1.3 Mise à jour du réseau 15
2.2 Régularisation entropique, Sinkhorn 18
2.2.1 Régularisation entropique 19
2.2.2 Algorithme de Sinkhorn 20

3 Des distances sur les distributions de probabilités **22**
3.1 La distance de p-Wasserstein 22
3.2 Les autres distances et divergences classiques 24
3.3 Comparaison des notions de convergence 25

4 Applications en apprentissage statistique **27**
4.1 Les réseaux antagonistes génératifs (GAN) 27
4.1.1 Fonctionnement des GAN 28
4.1.2 Apport du transport optimal 29
4.2 Adaptation de domaine 30
4.2.1 Motivation du problème de l'adaptation de domaine 30

4.2.2 Formalisation du problème 30
4.2.3 Apport du transport optimal 31

Bibliographie **34**

Introduction

Le problème du transport optimal peut être motivé par une situation bien française : chaque matin à Paris, quelques 1300 boulangeries doivent approvisionner en croissants les 2000 cafés de la ville. Il s'agit de le faire d'une manière qui a du sens. On veut transporter « optimalement » les croissants des boulangeries vers les cafés. Nous allons formaliser un peu tout cela :

Supposons que nous avons b_n la production en croissant de la n-ième boulangerie et c_n la demande en café du nième café. De plus, pour chaque couple boulangerie-café on a un coût $c_{\text{boulangerie} \to \text{café}}$ qui correspond au temps de transport pour déplacer un croissant de cette boulangerie à ce café.

Le problème du transport optimal consiste à trouver un plan de transport qui associe à chaque boulangerie les cafés qu'elle doit livrer, et qui minimise le temps de transport total de ces croissants dans Paris.

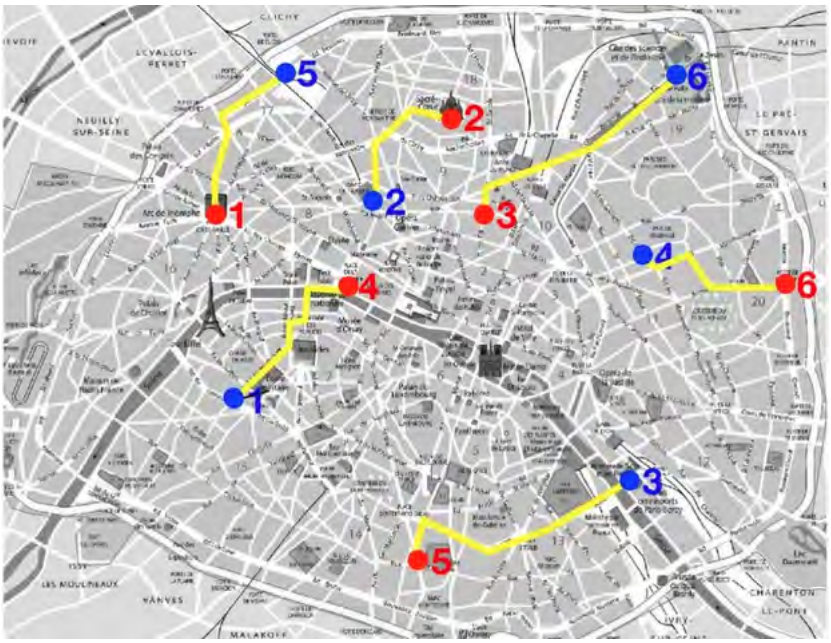


FIGURE 1 – Exemple de plan de transport

Issue d'un problème assez simple, la théorie du transport optimal a eu de nombreuses répercussions, que ce soit en mathématiques, en physique ou en économie. Porté par Monge au XVIII$^{\text{ème}}$ siècle, puis par Kantorovich au XX$^{\text{ème}}$ siècle, le domaine a connu un nouvel essor avec la contribution récente de Villani et les applications nouvellement trouvées dans le domaine de l'intelligence artificielle.

Ce mémoire s'appliquera à exposer les bases théoriques du problème du transport optimal et proposera un tour d'horizon des méthodes numériques qui permettent en pratique de le résoudre. Enfin nous présenterons deux exemples issus de la recherche de son application dans le domaine de l'apprentissage statistique.

Notations

- $\mathcal{X}$ et $\mathcal{Y}$ sont des espaces probabilisables.
- Si $\mathcal{X}$ est un espace probabilisable, on note $\mathrm{P}(\mathcal{X})$ l'ensemble des mesures de probabilité sur $\mathcal{X}$.
- $\mathrm{P}_f(\mathcal{X}) = \left\{ \sum_{x \in \mathcal{F}} \pi_x \, \delta_x, \mathcal{F} \subset \mathcal{X} \text{ fini }, (\pi_x)_{x \in \mathcal{F}} \in \mathbb{R}_+^{|\mathcal{F}|}, \sum_{x \in \mathcal{F}} \pi_x = 1 \right\}$ est l'ensemble des mesures de probabilité à support fini sur $\mathcal{X}$.
- $\mu \in \mathrm{P}(\mathcal{X})$ et $\nu \in \mathrm{P}(\mathcal{Y})$ sont des mesures de probabilité.
- $\Pi(\mu, \nu)$ est l'ensemble des mesures de probabilités sur $\mathcal{X} \times \mathcal{Y}$ dont les lois marginales respectives sont μ et ν, *i.e.* des couplages entre μ et ν.
- $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ est une fonction mesurable appelée fonction de coût.

Note. Dans ce mémoire, on supposera toujours que μ et ν sont des mesures à support fini (en particulier discrètes). Cependant, le problème du transport optimal existant entre deux mesures quelconques, il est intéressant de garder un formalisme aussi général que possible.

Chapitre 1

Le problème du transport optimal

1.1 Le problème selon Kantorovitch

1.1.1 Formulation générale

Soit $\mathcal{X}$ et $\mathcal{Y}$ deux espaces mesurables, et $\mu \in \mathrm{P}(\mathcal{X})$, $\nu \in \mathrm{P}(\mathcal{Y})$ deux mesures de probabilité sur $\mathcal{X}$ et $\mathcal{Y}$ respectivement.

Remarque 1.1.1. L'hypothèse importante est que les mesures soient de même masse totale finie (conservation de la masse lors du transport).

Le problème de Monge-Kantorovitch est une version relaxée et symétrisée par Kantorovitch du problème du transport optimal de Monge (dans le sens où l'on peut fractionner la masse d'un point de la distribution initiale). Il se formule de la manière suivante :

Problème 1.1.2 (Monge-Kantorovitch)**.**

$$\mathcal{L}_c(\mu,\nu) \overset{\text{déf.}}{=} \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\pi(x,y) = \inf_{\pi \in \Pi(\mu,\nu),\,(X,Y)\sim\pi} \mathbb{E}\left[c(X,Y)\right]$$

De manière plus compacte, on note :

$$\mathcal{L}_c(\mu,\nu) \overset{\text{déf.}}{=} \inf_{\pi \in \Pi(\mu,\nu)} \langle c,\, \pi \rangle$$

Le problème se comprend en interprétant $\mathrm{d}\pi(x,y)$ comme la proportion de la masse totale déplacée selon le couplage π entre x et y, dont le coût est $c(x,y)$. On veut donc bien minimiser le coût total du transport de μ à ν.

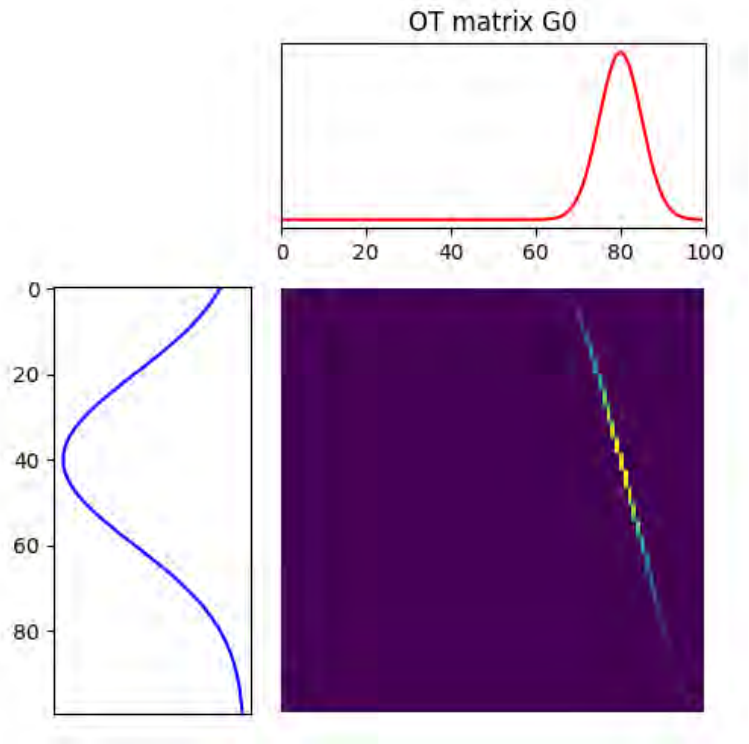


FIGURE 1.1 – Exemple de couplage entre deux gaussiennes

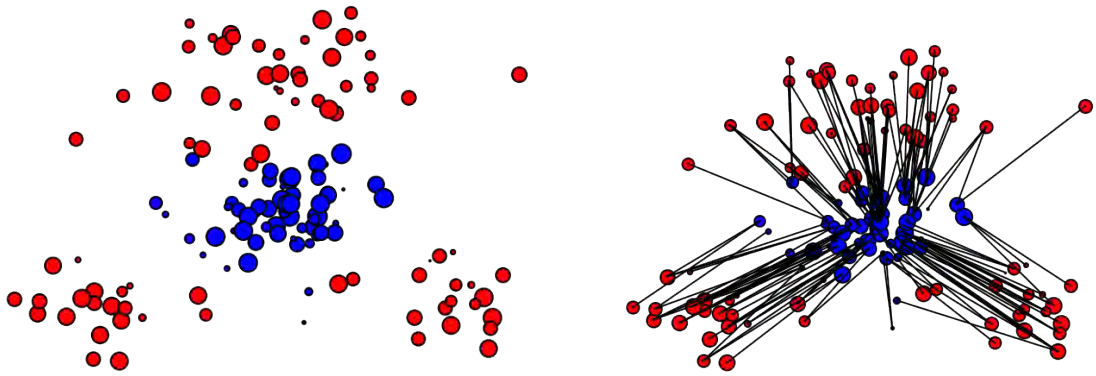


FIGURE 1.2 – Exemple de couplage entre deux distributions discrètes

1.1.2 Formulation discrète

Si $\mathcal{F} = \{x_1, \ldots, x_d\}$ est une partie finie de $\mathcal{X}$, et $\mu \in \mathrm{P}_f(\mathcal{X})$ est une mesure de probabilité à support fini inclus dans $\mathcal{F}$, on peut identifier μ à son vecteur de

probabilité via :

$$\begin{array}{rcl} \mathrm{P}_f(\mathcal{X}) & \longrightarrow & \Sigma_d = \{\, {}^t(\mathbf{a}_1, \ldots, \mathbf{a}_d) \in \mathbb{R}^d, \sum_{i=1}^{i=d} \mathbf{a}_i = 1\} \\ \mu & \longmapsto & (\mu(\{x_i\}))_{1 \leq i \leq d} \end{array} \tag{1.1}$$

Cette remarque est fondamentale car elle permet d'identifier le problème de Monge-Kantorovitch au problème d'optimisation linéaire (*i.e.* de minimisation d'une forme linéaire sur un polytope convexe en dimension finie) suivant :

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) \overset{\text{déf.}}{=} \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle \tag{1.2}$$

où $\mathbf{a} \in \Sigma_n$, $\mathbf{b} \in \Sigma_m$, $\mathbf{C} \in \mathcal{M}_{n,m}(\mathbb{R})$ et :

$$\mathbf{U}(\mathbf{a}, \mathbf{b}) \overset{\text{déf.}}{=} \left\{ \mathbf{P} \in \mathcal{M}_{n,m}(\mathbb{R}_+) \ : \ \mathbf{P}\mathbb{1}_m = \mathbf{a} \text{ et } {}^t\mathbf{P}\mathbb{1}_n = \mathbf{b} \right\} \tag{1.3}$$

Remarque 1.1.3. Cette traduction du problème dans sa version discrète est principalement introduite en raison de sa forme matricielle qui se prête bien aux méthodes d'optimisation linéaire de la section 1.2 et aux algorithmes de résolution qui seront présentés au chapitre 2.

1.1.3 Existence des solutions

Proposition 1.1.4 (Existence des solutions)**.** *Le problème de Monge-Kantorovitch admet des solutions,* i.e. *il existe au moins un couplage optimal π^* dans $\Pi(\mu, \nu)$ tel que :*

$$\mathcal{L}_c(\mu, \nu) = \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, \mathrm{d}\pi^*(x, y)$$

Démonstration du théorème 1.1.4. Dans le cas où μ et ν sont à support fini, on se ramène à prouver l'existence d'une solution au problème discret introduit à la sous-section 1.1.2. Or l'application :

$$\begin{array}{rcl} \mathbf{U}(\mathbf{a}, \mathbf{b}) & \longrightarrow & \mathbb{R} \\ P & \longmapsto & \langle C, P \rangle \end{array}$$

est linéaire donc continue et $\mathbf{U}(\mathbf{a}, \mathbf{b})$ est un fermé (comme intersection de nm demi-espaces fermés et de $n + m$ hyperplans affines), borné (pour la norme d'opérateur par exemple) donc compact ; elle atteint donc son minimum en un point $\mathbf{P}^*$ de $\mathbf{U}(\mathbf{a}, \mathbf{b})$. □

1.2 Dualité lagrangienne

Le déroulement naturel est alors d'introduire le lagrangien, afin de libérer les contraintes affines :

Définition 1.2.1. Le **lagrangien** $\Lambda_{\mathbf{C}}$ associé au problème $\mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$ est défini pour $\mathbf{P} \in \mathcal{M}(\mathbb{R}_+)$ et $(\mathbf{f}, \mathbf{g}) \in \mathbb{R}^m \times \mathbb{R}^n$ par :

$$\Lambda_{\mathbf{C}}(\mathbf{P}, \mathbf{f}, \mathbf{g}) = \langle \mathbf{C},\, \mathbf{P} \rangle + \langle \mathbf{a} - \mathbf{P}\mathbb{1}_m,\, \mathbf{f} \rangle + \langle \mathbf{b} - {}^t\mathbf{P}\mathbb{1}_n,\, \mathbf{g} \rangle$$

Alors on peut réécrire le problème discret sous la forme :

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in \mathcal{M}_{n,m}(\mathbb{R}_+)} \left(\max_{(\mathbf{f},\mathbf{g}) \in \mathbb{R}^m \times \mathbb{R}^n} \Lambda_{\mathbf{C}}(\mathbf{P}, \mathbf{f}, \mathbf{g}) \right) \tag{1.4}$$

Le problème dual apparaît lors de l'inversion entre min et max :

$$\begin{aligned} \widehat{\mathrm{L}}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) &\overset{\text{déf.}}{=} \max_{(\mathbf{f},\mathbf{g}) \in \mathbb{R}^m \times \mathbb{R}^n} \left(\min_{\mathbf{P} \in \mathcal{M}_{n,m}(\mathbb{R}_+)} \Lambda_{\mathbf{C}}(\mathbf{P}, \mathbf{f}, \mathbf{g}) \right) \\ &= \max_{(\mathbf{f},\mathbf{g}) \in \mathbb{R}^m \times \mathbb{R}^n} \langle \mathbf{f},\, \mathbf{a} \rangle + \langle \mathbf{g},\, \mathbf{b} \rangle + \min_{\mathbf{P} \in \mathcal{M}_{n,m}(\mathbb{R}_+)} \langle \mathbf{C} - \mathbf{f}\, {}^t\mathbb{1}_m - \mathbb{1}_n\, {}^t\mathbf{g},\, \mathbf{P} \rangle \end{aligned}$$

Or on a :

$$\min_{\mathbf{P} \in \mathcal{M}_{n,m}(\mathbb{R}_+)} \langle \mathbf{C} - \mathbf{f}\, {}^t\mathbb{1}_m - \mathbb{1}_n\, {}^t\mathbf{g},\, \mathbf{P} \rangle = \begin{cases} 0 & \text{si } \mathbf{C} - \mathbf{f}\, {}^t\mathbb{1}_m - \mathbb{1}_n\, {}^t\mathbf{g} \in \mathcal{M}_{n,m}(\mathbb{R}_+) \\ -\infty & \text{sinon} \end{cases}$$

Ainsi :

$$\widehat{\mathrm{L}}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \max_{(\mathbf{f},\mathbf{g}) \in \mathbf{R}(\mathbf{C})} \langle \mathbf{f},\, \mathbf{a} \rangle + \langle \mathbf{g},\, \mathbf{b} \rangle \tag{1.5}$$

où $\mathbf{R}(\mathbf{C}) = \{(\mathbf{f}, \mathbf{g}) \in \mathbb{R}^n \times \mathbb{R}^m, \forall\, (i,j) \in [\![1, m]\!] \times [\![1, n]\!], \mathbf{f}_i + \mathbf{g}_j \leq \mathbf{C}_{i,j}\}$.

Cette équation peut se réécrire dans le formalisme du problème initial dans le cas où les mesures sont à support fini :

Problème 1.2.2 (Monge-Kantorovitch, énoncé dual)**.**

$$\widehat{\mathcal{L}}_c(\mu, \nu) = \sup_{f,g \in \mathcal{R}(c)} \langle f,\, \mu \rangle + \langle g,\, \nu \rangle = \sup_{f,g \in \mathcal{R}(c)} \mathbb{E}_{X \sim \mu}\left[f(X)\right] + \mathbb{E}_{Y \sim \nu}\left[g(Y)\right]$$

Remarque 1.2.3. La régularité des potentiels f et g n'a que peu d'importance dans l'écriture ci-dessus pour des mesures à support fini. En revanche, l'énoncé dual général nécessite une hypothèse de continuité (ou le caractère 1-lipschitzien) des potentiels pour que les résultats ci-dessous puissent se généraliser.

Théorème 1.2.4 (Dualité faible, dualité forte)**.**

1. *(Dualité faible).*
$$\widehat{\mathcal{L}}_c(\mu, \nu) \leq \mathcal{L}_c(\mu, \nu)$$
2. *(Dualité forte).*
$$\widehat{\mathcal{L}}_c(\mu, \nu) = \mathcal{L}_c(\mu, \nu)$$

Démonstration. On exposera la preuve dans le cas où les mesures sont à support fini. La dualité générale est démontrée dans [Vil08].

1. (Dualité faible). Celle-ci découle simplement du fait que le plus grand des minima est inférieur au plus petit des maxima.
2. (Dualité forte). La dualité forte se démontre en suivant l'algorithme du simplexe, ou via le lemme de Farkas, qui se déduit du théorème de séparation point/convexe-fermé par des hyperplans. C'est un résultat général pour tous les problèmes d'optimisation linéaire admettant une solution.

□

Chapitre 2

Les algorithmes. Régularisation entropique

2.1 L'algorithme du simplexe

Le problème primal du transport optimal peut se réécrire sous cette forme :

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a},\mathbf{b}) = \min_{\substack{\mathbf{p}\in\mathbb{R}_+^{nm} \\ \mathbf{Ap}=\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix}}} {}^t\mathbf{cp}, \tag{2.1}$$

où p est l'énumération colonne par colonne de $\mathbf{P}$ et $\mathbf{A} = \begin{bmatrix} {}^t\mathbb{1}_n \otimes \mathbb{I}_m \\ \mathbb{I}_n \otimes {}^t\mathbb{1}_m \end{bmatrix} \in \mathcal{M}_{n+m,nm}(\mathbb{R})$ ($\otimes$ est le produit de Kronecker).

Ce problème, qui est un programme linéaire, est plus précisément appelé le problème du **min-cost flow**. Il existe des algorithmes efficaces pour le résoudre, y compris celui présenté dans cette section, le **network-simplex**. Avant de présenter l'algorithme, nous allons prouver quelques résultats intermédiaires qui seront très utiles pour expliquer son fonctionnement et sa correction.

Proposition 2.1.1. *Supposons que $\boldsymbol{P}^\star$ soit solution du problème primal, et $(\boldsymbol{f}^\star, \boldsymbol{g}^\star)$ solution du problème dual. On a alors :*

$$\forall (i,j) \in [\![1,n]\!] \times [\![1,m]\!], \boldsymbol{P}^\star_{i,j}(\boldsymbol{C}_{i,j} - \boldsymbol{f}^\star_i + \boldsymbol{g}^\star_j) = 0$$

Démonstration. Par la dualité forte on a que :

$$\begin{aligned}\langle \mathbf{P}^\star, \mathbf{C}\rangle &= \langle \mathbf{f}^\star, \mathbf{a}\rangle + \langle \mathbf{g}^\star, \mathbf{b}\rangle \\ &= \langle \mathbf{f}^\star, \mathbf{P}^\star \mathbb{1}_m\rangle + \langle \mathbf{g}^\star, {}^t\mathbf{P}^\star \mathbb{1}_n\rangle \\ &= \langle \mathbf{f}^\star {}^t\mathbb{1}_m, \mathbf{P}^\star\rangle + \langle \mathbb{1}_n {}^t\mathbf{g}^\star, \mathbf{P}^\star\rangle \\ &= \langle \mathbf{f}^\star {}^t\mathbb{1}_m + \mathbb{1}_n {}^t\mathbf{g}^\star, \mathbf{P}^\star\rangle\end{aligned}$$

d'où :

$$\langle \mathbf{P}^\star, \mathbf{C} - \mathbf{f}^\star \oplus \mathbf{g}^\star\rangle = 0$$

où l'on note $\mathbf{f}^\star \oplus \mathbf{g}^\star = (\mathbf{f}_i^\star + \mathbf{g}_j^\star)_{(i,j)\in[\![1,n]\!]\times[\![1,m]\!]}$. □

Le résultat réciproque est également vrai et va particulièrement nous intéresser dans le cadre du network-simplex.

Définition 2.1.2 (Paire complémentaire)**.** Une matrice $\mathbf{P} \in \mathcal{M}_n(\mathbb{R})$ et un couple $(\mathbf{f}, \mathbf{g}) \in \mathbb{R}^n \times \mathbb{R}^m$ forment une **paire complémentaire** par rapport à $\mathbf{C}$ si, et seulement si, pour tout $(i, j) \in [\![1, n]\!] \times [\![1, m]\!]$ tel que $\mathbf{P}_{i,j} > 0$, on a : $\mathbf{C}_{i,j} = \mathbf{f}_i + \mathbf{g}_j$.

Proposition 2.1.3 (Optimalité d'une paire complémentaire)**.** *Si $\boldsymbol{P}$ et $(\boldsymbol{f}, \boldsymbol{g})$ sont une paire complémentaire et respectivement des solution admissibles du problème primal et dual, alors elles sont des solutions optimales.*

Démonstration. Par la dualité faible, on a :

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) \leq \langle \mathbf{P}, \mathbf{C}\rangle = \langle \mathbf{P}, \mathbf{f} \oplus \mathbf{g}\rangle = \langle \mathbf{a}, \mathbf{f}\rangle + \langle \mathbf{b}, \mathbf{g}\rangle \leq \mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$$

Et donc :

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

□

Un programme linéaire avec un ensemble de contraintes borné atteint sa solution optimale à un point extrémal du polyèdre convexe des solutions admissibles. Une manière de trouver une solution optimale est donc de visiter ces points particuliers de $\mathbf{U}(\mathbf{a}, \mathbf{b})$, et de vérifier à chaque fois s'il s'agit d'une solution optimale. Nous allons montrer quelques propriétés sur les plans de transports associés aux points extrémaux de $\mathbf{U}(\mathbf{a}, \mathbf{b})$.

Nous allons représenter notre problème sous la forme d'un graphe biparti. On considère $V = (1, 2, ..., n)$ et $V' = (1', 2', ..., m')$ deux ensembles de n et m sommets respectivement, et E les nm arêtes reliant chaque sommet de V à V'. On définit alors $G = (V \cup V', E)$. A chaque arête (i, j) on associe le coût $\mathbf{C}_{i,j}$. On peut donc reformuler le problème en un problème de flot sur le graphe G. Il s'agit de trouver

un flot sur le graphe qui minimise le coup total des arêtes utilisées et qui vérifie deux propriétés :

- le flot sortant du sommet k est égal à $\mathbf{a}_k$;
- le flot entrant dans un sommet l' est égal à $\mathbf{b}_{l'}$.

On note $G(\mathbf{P})$ un graphe correspondant à un plan de transport $\mathbf{P}$ et où on n'a gardé que les arêtes de flot non nul. Formellement, $S(\mathbf{P}) = \{(i,j) \in E, \mathbf{P}_{i,j} > 0\}$ et $G(\mathbf{P}) = (V \cup V', S(\mathbf{P}))$.

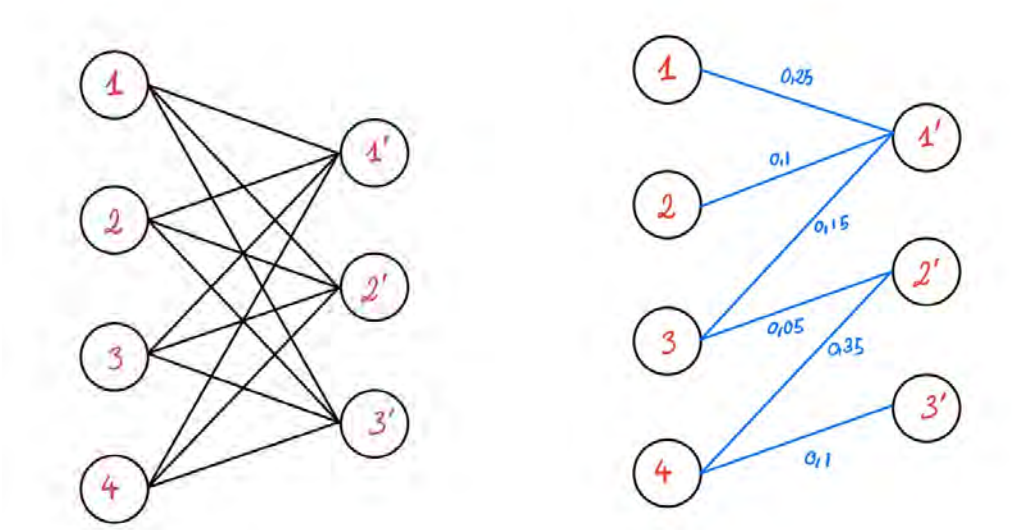


FIGURE 2.1 – Exemple du graphe G pour $\mathbf{a} = [0.35, 0.2, 0.45]$ et $\mathbf{b} = [0.5, 0.4, 0.1]$ et d'un plan de transport admissible.

Proposition 2.1.4. *Soit $\boldsymbol{P}$ un point extrémal de $\boldsymbol{U}(\boldsymbol{a}, \boldsymbol{b})$. Alors $G(\boldsymbol{P})$ est sans cycle et en particulier $\boldsymbol{P}$ a au plus $n + m - 1$ entrées non nulles.*

Démonstration. On suppose que $\mathbf{P}$ est un point extrémal de $\mathbf{U}(\mathbf{a}, \mathbf{b})$ et que $G(\mathbf{P}) = (V \cup V', S(\mathbf{P}))$ possède un cycle et on va montrer le résultat par l'absurde. On suppose donc qu'il existe $i_1, ..., i_k \in [\![1, n]\!]$ et $j_i, ..., j_k \in [\![1, m]\!]$ tels que :

$$H = \{(i_1, j_1'), (j_1', i_2), (i_2, j_2'), ..., (j_k', i_1)\} \subset S(\mathbf{P})$$

(Notez que le graphe G n'est pas dirigé et que les arêtes sont écrites de cette manière pour faciliter la lecture de la preuve). On va maintenant construire $\mathbf{Q}$ et $\mathbf{R}$ tels que $P = \frac{\mathbf{Q}+\mathbf{R}}{2}$, et $\mathbf{Q}, \mathbf{R} \neq \mathbf{P}$. Si on réussit une telle construction cela voudra dire que $\mathbf{P}$ n'est pas un point extrémal de $\mathbf{U}(\mathbf{a}, \mathbf{b})$ et cela conclura notre preuve par l'absurde.

On pose $\varepsilon < \min_{(i,j)\in S(\mathbf{P})} \mathbf{P}_{i,j}$, et $\mathbf{E}$ une matrice telle que :

$$\mathbf{E}_{i,j} = \begin{cases} \varepsilon & \text{si } (i,j') \in H \\ -\varepsilon & \text{si } (j',i) \in H \\ 0 & \text{sinon} \end{cases}$$

On pose maintenant $\mathbf{Q} = \mathbf{P} + \mathbf{E}$ et $\mathbf{R} = \mathbf{P} - \mathbf{E}$. On a choisi ε suffisamment petit pour que $\mathbf{R}$ n'ait pas d'élément négatif. Par construction, chaque colonne (respectivement ligne) de $\mathbf{E}$ comporte soit uniquement des zéros, soit un unique indice de valeur ε et un unique indice de valeur $-\varepsilon$. On a donc $\mathbf{E}\mathbb{1}_m = \mathbb{0}_n$ et ${}^t\mathbf{E}\mathbb{1}_n = \mathbb{0}_m$, et donc $\mathbf{Q}$ et $\mathbf{R}$ sont des solutions admissibles. □

Remarque 2.1.5. Comme $G(\mathbf{P})$ est sans cycle, c'est soit un arbre, c'est-à-dire un graphe sans cycle connexe, soit une forêt, c'est à dire une union d'arbres.

L'idée générale de l'algorithme est assez classique. On sélectionne un point extrémal arbitraire de $\mathbf{U}(\mathbf{a}, \mathbf{b})$. Si c'est une solution optimale, on s'arrête. Sinon on sélectionne un autre point extrémal de manière à ce qu'il ait un coût plus faible. On continue l'opération jusqu'à tomber sur une solution optimale. Nous allons montrer pas-à-pas que cet algorithme est en effet implémentable.

2.1.1 Sélectionner un point de départ pour notre algorithme

Une manière d'obtenir un point extrémal de $\mathbf{U}(\mathbf{a}, \mathbf{b})$ est d'utiliser la **méthode du Nord-Ouest**, qui permet d'obtenir un tel point en $n + m$ opérations. Nous la

décrivons dans ce paragraphe.

Algorithme 1 : Règle du Nord-Ouest

$i \leftarrow 1$
$j \leftarrow 1$
$r \leftarrow \mathbf{a}_1$
$c \leftarrow \mathbf{b}_1$
while $i \leq n$ *et* $j \leq m$ **do**
  $t \leftarrow \min(r, c)$
  $\mathbf{P}_{i,j} \leftarrow t$
  $r \leftarrow r - t$
  $c \leftarrow s - t$
  if $r = 0$ **then**
    $i \leftarrow i + 1$
    $r \leftarrow \mathbf{a}_i$
  end
  if $c = 0$ **then**
    $j \leftarrow j + 1$
    $c \leftarrow \mathbf{b}_j$
  end
end

Exemple 2.1.6. $\mathbf{a} = [0.4, 0.3, 0.3]$ and $\mathbf{b} = [0.5, 0.2, 0.3]$:

$$\begin{bmatrix} \bullet & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0.4 & 0 & 0 \\ \bullet & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0.4 & 0 & 0 \\ 0.1 & \bullet & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} 0.4 & 0 & 0 \\ 0.1 & 0.2 & 0 \\ 0 & 0 & \bullet \end{bmatrix} \rightarrow \begin{bmatrix} 0.4 & 0 & 0 \\ 0.1 & 0.2 & 0 \\ 0 & 0 & 0.3 \end{bmatrix}$$

2.1.2 Vérifier qu'une solution est ou non optimale

On a montré à la proposition 2.1.3 une méthode pour vérifier si une solution $\mathbf{P}$ est ou non optimale. Il suffit de prendre sa paire duale $(\mathbf{f}, \mathbf{g})$ et de vérifier s'il s'agit d'un couple admissible du problème dual.

Étant donné un plan de transport admissible $\mathbf{P}$, on peut trouver un paire complémentaire en résolvant : $\forall\,(i,j) \in [\![1,n]\!] \times [\![1,m]\!], \mathbf{f}_i + \mathbf{g}_j = \mathbf{C}_{i,j}$.

2.1.3 Mise à jour du réseau

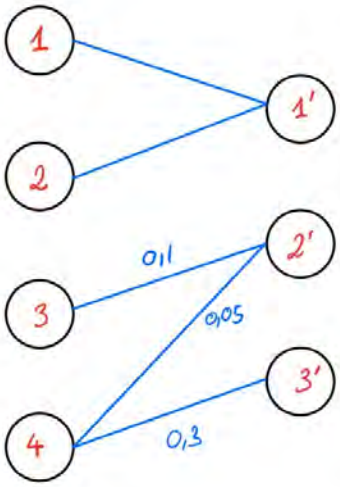
Si la solution $\mathbf{P}$ n'est pas optimale, c'est qu'il existe (i,j) tel que $\mathbf{f}_i + \mathbf{g}_j < \mathbf{C}_{i,j}$.

On ajoute alors au graphe $G(\mathbf{P})$ l'arête (i, j). Deux cas de figures peuvent alors se présenter :

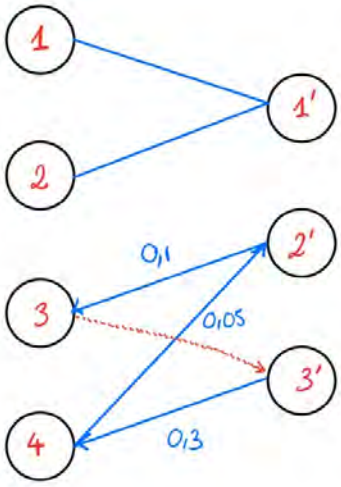
1. L'arête (i, j) relie deux sous-arbres de la forêt $G(\mathbf{P})$. $G(\mathbf{P})$ est donc toujours une forêt, et donc $\mathbf{P}$ ne change pas. On peut utiliser la procédure 2.1.2 pour calculer à nouveau la paire complémentaire $(\mathbf{f}, \mathbf{g})$.
2. Le graphe $G(\mathbf{P})$ a maintenant un cycle, que l'on note :

   $$\{(i_1, j'_1), (j'_1, i_2), ..., (j'_l, i_1)\} \text{ où } (i_1, j'_1) = (1, j)$$

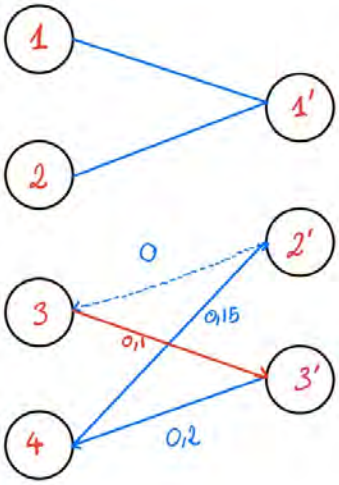
   On appelle arêtes « positives » (respectivement « négatives ») les arêtes de ce cycle de la forme (i_k, j'_k) (respectivement (j'_k, i_{k+1})). On définit ε comme étant le flot minimal des arêtes « négatives » du cycle. Par la suite, on diminue le flot de toutes les arêtes négatives du cycle de ε, et on augmente de la même quantité les autres arêtes du cycle. Cela permet d'enlever l'arête négative de flot minimal et de rompre le cycle. On obtient alors la solution mise à jour $\tilde{\mathbf{P}}$ et on peut calculer la nouvelle paire duale complémentaire $(\mathbf{f}, \mathbf{g})$.



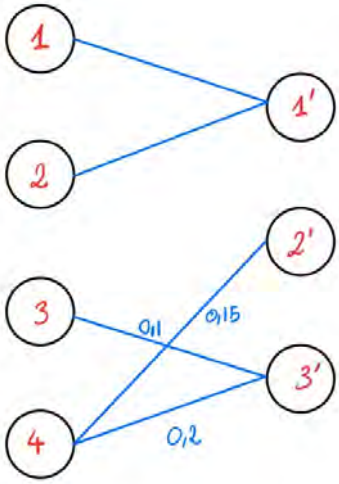
(a) $G(\mathbf{P})$ avant la mise à jour du réseau



(b) L'ajout de l'arête $(3, 3')$ crée un cycle



(c) Mise à jour des poids des arêtes



(d) $G(\tilde{\mathbf{P}})$

FIGURE 2.2 – Exemple de mise à jour du **network-simplex** dans le cas où l'ajout d'une arête entraîne la création d'un cycle.

Théorème 2.1.7. *La procédure décrite dans le paragraphe 2.1.3 améliore le coût de la solution $\boldsymbol{P}$. C'est-à-dire que si $\boldsymbol{P}$ est mis à jour en $\tilde{\boldsymbol{P}}$, on a $\langle \boldsymbol{P}, \boldsymbol{C} \rangle \geq \langle \tilde{\boldsymbol{P}}, C \rangle$.*

Démonstration. Le cas intéressant ici est de voir ce qu'il se passe lorsque qu'un cycle est créé au cours de la mise à jour du **network-simplex**. Dans ce cas, $\mathbf{P}$ est mis à jour en $\tilde{\mathbf{P}}$ et on a cette équation :

$$\langle \tilde{\mathbf{P}}, \mathbf{C} \rangle - \langle \mathbf{P}, \mathbf{C} \rangle = \varepsilon \left(\sum_{k=1}^{l} \mathbf{C}_{i_k, j_k} - \sum_{k=1}^{l} \mathbf{C}_{i_{k+1}, j_k} \right)$$

Cela correspond à l'augmentation de coût correspondant aux arêtes où on a rajouté ε aux poids moins la diminution correspondant aux arêtes négatives où on a enlevé ε aux poids. Soit $(\mathbf{f}, \mathbf{g})$ la paire complémentaire associée à $\mathbf{P}$. Elle vérifie $\forall\, (v, w), \mathbf{f}_v + \mathbf{g}_w = C_{vw}$. Donc on peut réécrire :

$$\begin{aligned} \sum_{k=1}^{l} \mathbf{C}_{i_k, j_k} - \sum_{k=1}^{l} \mathbf{C}_{i_{k+1}, j_k} &= \mathbf{C}_{i,j} + \sum_{k=2}^{l} \mathbf{f}_{i_k} + \mathbf{g}_{j_k} - \sum_{k=1}^{l} \mathbf{f}_{i_{k+1}} + \mathbf{g}_{j_k} \\ &= \mathbf{C}_{i,j} - (\mathbf{f}_i + \mathbf{g}_j). \end{aligned}$$

Ce dernier terme est négatif car i et j ont été choisis car $\mathbf{f}_i + \mathbf{g}_j < \mathbf{C}_{i,j}$, ce qui prouve le résultat. □

Finalement l'algorithme du **network-simplex** peut être résumé ainsi :

Algorithme 2 : Network-Simplex

$P \leftarrow NW(\mathbf{a}, \mathbf{b})$
$G \leftarrow G(\mathbf{P})$
$(\mathbf{f}, \mathbf{g}) \leftarrow calcul\_paire\_complementaire(\mathbf{P})$
while ***P*** *et* $(\boldsymbol{f}, \boldsymbol{g})$ *ne sont pas des solutions admissibles* **do**
  Faire une étape de mise à jour du network-simplex décrite en 2.1.3
end

Il a été prouvé dans [Tar97] que cet algorithme a une complexité polynomiale et plus particulièrement une complexité de :

$$O\left((n+m)nm \log(n+m) \log\left((n+m) \|\mathbf{C}\|_\infty \right) \right)$$

En pratique, cet algorithme est très efficace.

2.2 Régularisation entropique, Sinkhorn

Bien que l'algorithme décrit dans la section précédente soit très efficace en pratique, lorsque la dimension de $\mathbf{a}$ et $\mathbf{b}$ dépasse quelques centaines, le coût d'exé-

cution du network-simplex devient prohibitif. C'est ce qui a retardé pendant longtemps l'utilisation du transport optimal dans des contextes d'apprentissage avec des données assez larges. Cependant, en 2013, [Cut13] propose une régularisation du problème du transport optimal qui permet d'approximer un plan de transport optimal très rapidement.

2.2.1 Régularisation entropique

Définition 2.2.1. On définit l'entropie discrète d'une matrice de couplage $\mathbf{P}$ par :

$$\mathbf{H}(\mathbf{P}) \overset{\text{déf.}}{=} -\sum_{i,j} \mathbf{P}_{i,j}(\log(\mathbf{P}_{i,j}) - 1)$$

On va maintenant pouvoir définir le problème du transport optimal régularisé.

Problème 2.2.2. *Le problème du transport optimal régularisé correspond à résoudre :*

$$\mathrm{L}^{\varepsilon}_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b}) \overset{\text{déf.}}{=} \min_{\boldsymbol{P} \in \boldsymbol{U}(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P},\ \boldsymbol{C} \rangle - \varepsilon \boldsymbol{H}(\boldsymbol{P})$$

Le terme $\mathbf{H}(\mathbf{P})$ est donc utilisé ici comme un terme de régularisation pour obtenir des solutions approchées du problème du transport optimal.

Remarque 2.2.3. On peut vérifier qu'on a en particulier :

1. $\mathrm{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) \overset{\varepsilon \to 0}{\longrightarrow} \mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$
2. $\mathbf{P}_{\varepsilon} \overset{\varepsilon \to \infty}{\longrightarrow} \mathbf{a} \otimes \mathbf{b} = \mathbf{a}\, {}^t\mathbf{b} = (\mathbf{a}_i\, \mathbf{b}_j)_{(i,j) \in [\![1,n]\!] \times [\![1,m]\!]}$

On s'aperçoit également empiriquement que les solutions associées au problème régularisé sont plus diffuses, comme on peut le voir avec la figure 2.3.

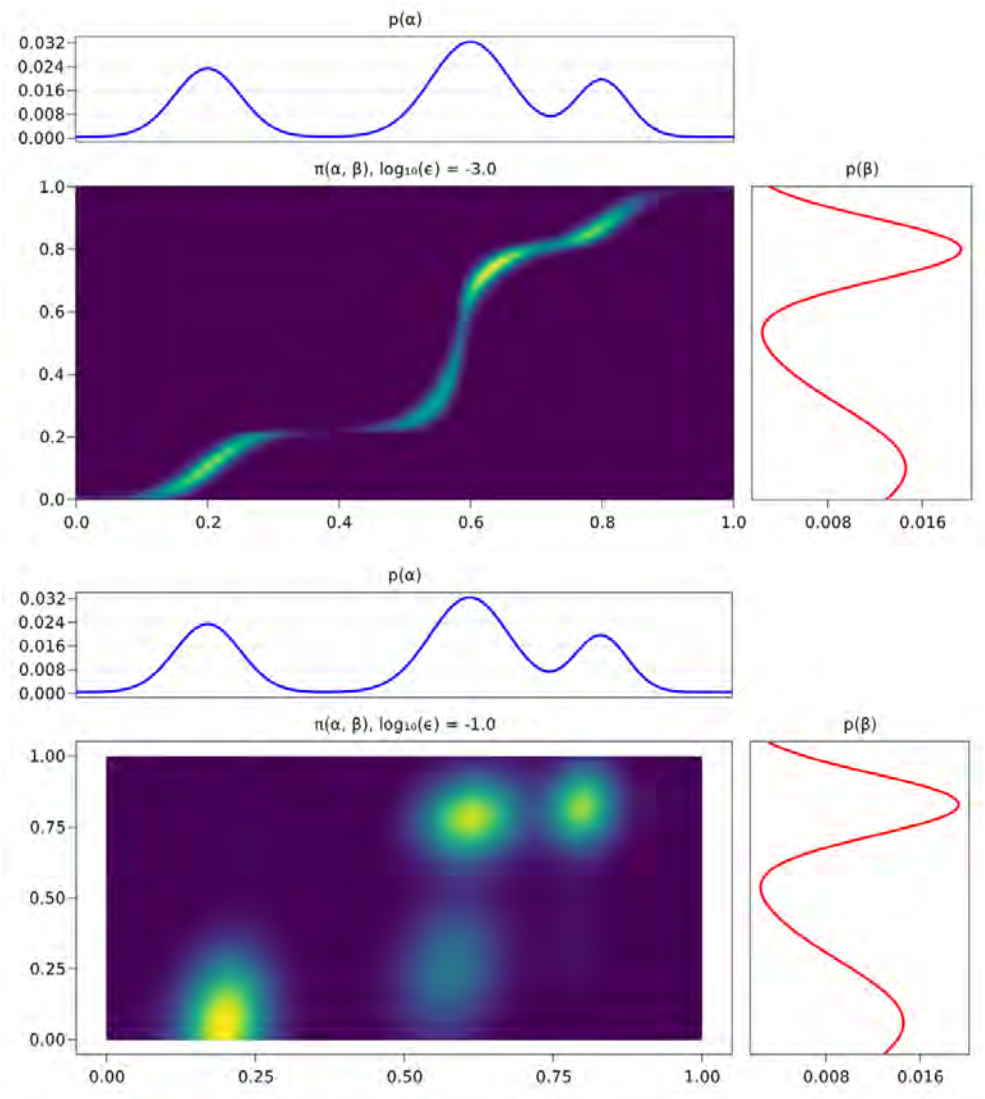


FIGURE 2.3 – Exemple de solutions du problème régularisé pour un ε relativement petit (en haut) et relativement grand (en bas).

2.2.2 Algorithme de Sinkhorn

Proposition 2.2.4. *La solution au transport optimal régularisé à la forme :*

$$\forall\,(i,j) \in [\![1,n]\!] \times [\![1,m]\!], \quad \boldsymbol{P}_{i,j} = \boldsymbol{u}_i \boldsymbol{K}_{i,j} \boldsymbol{v}_j$$

où K est le kernel de Gibbs, $\boldsymbol{K}_{i,j} \overset{\text{déf.}}{=} \exp\left(-\frac{\boldsymbol{C}_{i,j}}{\varepsilon}\right)$, et où $\boldsymbol{u}$ et $\boldsymbol{v}$ sont des inconnues.

Démonstration. On écrit le lagrangien $\Lambda_{\mathbf{C}}^{\varepsilon}$ associé au problème $\mathrm{L}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b})$

$$\Lambda_{\mathbf{C}}^{\varepsilon}(\mathbf{P}, \mathbf{f}, \mathbf{g}) = \langle \mathbf{P},\, \mathbf{C} \rangle - \varepsilon \mathbf{H}(\mathbf{P}) - \langle \mathbf{f},\, \mathbf{P}\mathbb{1}_m - \mathbf{a} \rangle - \langle \mathbf{g},\, {}^t\mathbf{P}\mathbb{1}_n - \mathbf{b} \rangle$$

Lorsque le minimum est atteint la dérivée s'annule et on a donc :

$$\frac{\partial \mathcal{E}(\mathbf{P}, \mathbf{f}, \mathbf{g})}{\partial \mathbf{P}_{i,j}} = \mathbf{C}_{i,j} + \varepsilon \log(\mathbf{P}_{i,j}) - \mathbf{f}_i - \mathbf{g}_j = 0$$

D'où, si P est une solution optimale on a $\mathbf{P}_{i,j} = e^{\mathbf{f}_i/\varepsilon} e^{-\mathbf{C}_{i,j}/\varepsilon} e^{\mathbf{g}_j/\varepsilon}$ ce qui correspond à l'expression recherchée. □

On peut donc réécrire le transport optimal régularisé sous forme matricielle ainsi :

$$\mathbf{P} = \mathrm{diag}(\mathbf{u})\mathbf{K}\,\mathrm{diag}(\mathbf{v})$$

avec les conditions sur $\mathbf{u}$ et $\mathbf{v}$ suivantes et dues à la conservation de la masse :

$\mathbf{u} \odot (\mathbf{K}\mathbf{v}) = \mathbf{a}$ et $\mathbf{v} \odot ({}^t\mathbf{K}\mathbf{u}) = \mathbf{b}$ et où $\odot$ correspond à la multiplication vectorielle terme à terme.

Il se trouve que le problème 2.2.2 réécrit sous cette forme est une instance d'un problème numérique bien connu appelé **matrix scaling** et dont il existe un algorithme efficace que nous allons décrire ci-dessous.

Algorithme 3 : Algorithme de Sinkhorn

$\mathbf{u} \leftarrow \mathbb{1}_n$
$\mathbf{v} \leftarrow \mathbb{1}_m$
$\mathbf{P} \leftarrow \mathrm{diag}(\mathbf{u})\mathbf{K}\,\mathrm{diag}(\mathbf{v})$
while $\boldsymbol{P}$ *change* **do**
  $\mathbf{u} \leftarrow \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}}$
  $\mathbf{v} \leftarrow \frac{\mathbf{b}}{{}^t\mathbf{K}\mathbf{u}}$
  $\mathbf{P} \leftarrow \mathrm{diag}(\mathbf{u})\mathbf{K}\,\mathrm{diag}(\mathbf{v})$
end

Il est prouvé dans [Cut13] que la complexité de cet algorithme est en $O(n^3 log(n))$ avec une complexité empirique en $O(n^2)$. L'algorithme du Sinkhorn calculant essentiellement des multiplications matricielles est également très parallélisable (pour par exemple calculer plusieurs solutions au problème 2.2.2 en même temps) et peut notamment être exécuté sur GPU, ce qui rend cet algorithme bien plus efficace dans des contextes d'apprentissage, où il y a beaucoup de données à traiter.

La convergence de l'algorithme de Sinkhorn est assurée comme il a été montré par Franklin dans [FL89].

Chapitre 3

Des distances sur les distributions de probabilités

Dans ce chapitre, on supposera que $\mathcal{X} = \mathcal{Y}$ est un espace métrique muni d'une distance d et de sa tribu de Borel $\mathcal{B}(\mathcal{X})$. On va montrer qu'alors la distance d sur $\mathcal{X}$ induit, via le transport optimal, une famille de distances sur un sous-espace de $\mathrm{P}(\mathcal{X})$ contenant les mesures à support fini (et égal à l'espace tout entier si l'on suppose en outre $\mathcal{X}$ compact).

3.1 La distance de p-Wasserstein

Définition 3.1.1. On définit, si $p \geq 1$, la **distance de p-Wasserstein** par :

$$\mathcal{W}_p(\mu, \nu) = (\mathcal{L}_{d^p}(\mu, \nu))^{\frac{1}{p}} \in [0, +\infty]$$

Proposition 3.1.2. *Si $\mathcal{X}$ est séparable complet, la distance de p-Wasserstein est une distance sur :*

$$\mathrm{P}_p(\mathcal{X}) = \{\mu \in \mathrm{P}(\mathcal{X}), \exists\, x_0 \in \mathcal{X}, \langle d(x_0, \cdot)^p,\, \mu\rangle < +\infty\}$$

Démonstration. Présentons la preuve proposée dans [Vil08]. Montrons d'abord que $\mathcal{W}_p$ vérifie les axiomes d'une distance, nous montrerons ensuite qu'elle est à valeurs réelles sur $\mathrm{P}_p(\mathcal{X})$.

Soit μ, ν, ξ des mesures de probabilité sur $\mathcal{X}$.

1. (Symétrie). Elle découle de celle du problème de Monge-Kantorovitch.
2. (Séparation). Notons $\pi \in \mathit{\Pi}(\mu, \mu)$ le couplage défini par :

$$\pi(A \times B) = \mu(A \cap B) \text{ si } A, B \in \mathcal{B}(\mathcal{X})$$

de sorte que si $(X, Y) \sim \pi$, on a $X = Y$ p.s. d'où $\mathbb{E}[d(X, Y)] = 0$ et π est optimal : $\mathcal{W}_1(\mu, \mu) = 0$. Réciproquement, si $\mathcal{W}_p(\mu, \nu) = 0$, soit (X, Y) un couplage optimal de (μ, ν) (dont l'existence est garantie par le théorème 1.1.4). Alors $d(X, Y) = 0$ p.s., *i.e.* $X = Y$ p.s. (par séparation de d), et en particulier $X \sim Y$. D'où $\mu = \nu$.

3. (Inégalité triangulaire). Commençons par fixer un couplage optimal (X, Y) (et (Y, Z)) de (μ, ν) (respectivement (ν, ξ)), on utilise ensuite le lemme gluant, prouvé dans [Vil03] :

   Lemme 3.1.3 (Lemme gluant)**.** *Soit $\mathcal{X}_1$, $\mathcal{X}_2$, $\mathcal{X}_3$ des espaces polonais (métriques, complets, séparables), et μ_1, μ_2, μ_3 des mesures de probabilité sur ces espaces respectifs. Si $\pi_{12} \in \Pi(\mu_1, \mu_2)$ et $\pi_{23} \in \Pi(\mu_2, \mu_3)$ sont deux couplages, alors il existe une mesure de probabilité $\pi \in \mathrm{P}(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3)$ ayant pour marges π_{12} sur $\mathcal{X}_1 \times \mathcal{X}_2$ et π_{23} sur $\mathcal{X}_2 \times \mathcal{X}_3$.*

   Par le lemme, on dispose d'un triplet $(X', Y', Z') \in \mathrm{P}(\mathcal{X}^3)$ tel que (X', Y') et (Y', Z') soient égales en loi respectivement à (X, Y) et (Y, Z) ; en particulier, (X', Z') est un couplage de (μ, ξ). D'où :

$$\begin{aligned}
\mathcal{W}_p(\mu, \xi) &\leq \mathbb{E}\left[d(X', Z')^p\right]^{\frac{1}{p}} \\
&\leq \mathbb{E}\left[(d(X', Y') + d(Y', Z'))^p\right]^{\frac{1}{p}} \quad \text{(inégalité triangulaire)} \\
&\leq \left(\mathbb{E}\left[d(X', Y')^p\right] + \mathbb{E}\left[d(Y', Z')^p\right]\right)^{\frac{1}{p}} \quad \text{(inégalité de Minkovski)} \\
\mathcal{W}_p(\mu, \xi) &\leq \mathcal{W}_p(\mu, \nu) + \mathcal{W}_p(\nu, \xi)
\end{aligned}$$

   C'est l'inégalité triangulaire attendue.

Enfin, montrons que W_p prend des valeurs finies sur $\mathrm{P}_p(\mathcal{X})$. Si μ et ν sont dans $\mathbf{P}_p(\mathcal{X})$, et $\pi \in \Pi(\mu, \nu)$, alors l'inégalité (découlant par exemple du binôme de Newton) :

$$d(x, y) \leq 2^{p-1}(d(x, x_0)^p + d(x_0, y)^p)$$

montre qu'il suffit que $d(\cdot, x_0)^p$ soit μ-intégrable et $d(x_0, \cdot)^p$ soit ν-intégrable pour que d^p soit π-intégrable, et de plus que la définition de $\mathrm{P}_p(\mathcal{X})$ ne dépend pas du choix de $x_0 \in \mathcal{X}$. □

Exemple 3.1.4. On a immédiatement que la distance de p-Wasserstein entre deux mesures de Dirac est donnée par :

$$\mathcal{W}_p(\delta_x, \delta_y) = d(x, y) \text{ si } (x, y) \in \mathcal{X}^2$$

3.2 Les autres distances et divergences classiques

L'idée de quantifier la dissimilarité entre deux distributions n'est pas nouvelle, et de nombreuses divergences ont été étudiées, notamment la divergence de Kullback-Leibler, qui s'interprète comme l'entropie relative des deux distributions.

Définition 3.2.1. Une **divergence** sur un ensemble $\mathcal{E}$ est une application d de $\mathcal{E}^2$ dans $\mathbb{R}_+$ séparant les points :

$$\forall (x, y) \in \mathcal{E}^2, x = y \iff d(x, y) = 0$$

Définition 3.2.2.

1. (Distance en variation totale).

$$\delta(\mu, \nu) = \sup_{A \in \mathcal{B}(\mathcal{X})} |\mu(A) - \nu(A)| \tag{3.1}$$

2. (Divergence de Kullback-Leibler).

$$\mathrm{KL}(\mu, \nu) = \int_{\mathcal{X}} \log\left(\frac{\mu(x)}{\nu(x)}\right) \mu(x)\, \mathrm{d}\rho(x) \tag{3.2}$$

   où μ et ν sont absolument continues par rapport à une mesure $\rho \in \mathrm{P}(\mathcal{X})$.

3. (Divergence de Jensen-Shannon).

$$\mathrm{JS}(\mu, \nu) = \mathrm{KL}(\mu, \xi) + \mathrm{KL}(\nu, \xi) \text{ où } \xi = \frac{\mu + \nu}{2} \tag{3.3}$$

Exemple 3.2.3. Prenons $\mathcal{X} = \mathbb{R}^2$ et $Y \sim \mathcal{U}([0, 1])$. Considérons la famille de mesures de probabilité $(\mu_\theta)_{\theta \in \mathbb{R}}$ où μ_θ est la loi de la variable aléatoire (θ, Y) (*i.e.* la distribution uniforme sur le segment vertical d'abscisse θ). Alors on a :

- $\mathcal{W}_1(\mu_\theta, \mu_0) = |\theta|$
- $\delta(\mu_\theta, \mu_0) = \delta_\theta$ (symbole de Kronecker)
- $\mathrm{KL}(\mu_\theta, \mu_0) = +\infty \times \delta_\theta$
- $\mathrm{JS}(\mu_\theta, \mu_0 = \log 2 \times \delta_\theta$

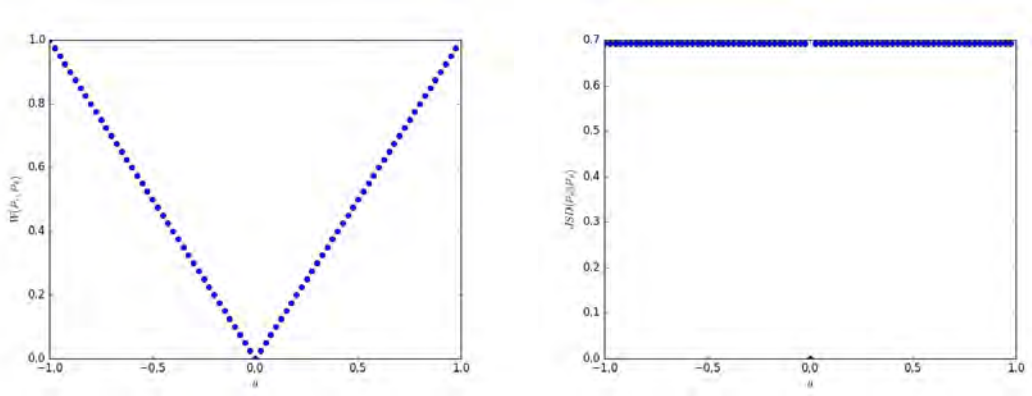


FIGURE 3.1 – Tracés de $\mathcal{W}_1(\mu_\theta, \mu_0)$ (à gauche) et $\mathrm{JS}(\mu_\theta, \mu_0)$ (à droite) en fonction de θ, illustrant la différence de régularité et de finesse de convergence entre la distance de Wasserstein et la divergence de Jensen-Shannon.

3.3 Comparaison des notions de convergence

Le résultat suivant a été démontré dans [ACB17] afin de comparer les notions de convergence induites par les différentes distances et divergences :

Théorème 3.3.1. *Supposons que $\mathcal{X}$ soit compact. Soit μ et $(\mu_n)_{n\in\mathbb{N}}$ dans $\mathrm{P}(\mathcal{X})$. Alors :*

1. $\mathrm{KL}(\mu_n, \mu) \to 0$ *ou* $\mathrm{KL}(\mu, \mu_n) \to 0 \implies \delta(\mu_n, \mu) \to 0$
2. $\delta(\mu_n, \mu) \to 0 \iff \mathrm{JS}(\mu_n, \mu) \to 0$
3. $\delta(\mu_n, \mu) \implies \mathcal{W}_1(\mu_n, \mu) \to 0$
4. $\mathcal{W}_1(\mu_n, \mu) \to 0$ *si, et seulement, si μ_n converge étroitement vers μ.*

Démonstration.

1. Le premier résultat découle de l'inégalité de Pinsker :

$$\delta(\mu_n, \mu) \leq \sqrt{\frac{1}{2}\,\mathrm{KL}(\mu_n, \mu)} \to 0 \text{ et } \delta(\mu, \mu_n) \leq \sqrt{\frac{1}{2}\,\mathrm{KL}(\mu, \mu_n)} \to 0 \quad (3.4)$$

2. Ce résultat est admis, la preuve (calculatoire) est à retrouver en annexe de [ACB17].
4. La démonstration est celle de Villani dans [Vil08].
3. Par le théorème de représentation de Riesz, $(\mathrm{P}(\mathcal{X}), \delta)$ est isométrique à une sous-partie de l'espace dual des fonctions continues sur $\mathcal{X}$. Sa topologie est donc plus fine que la topologie faible $\star$, d'où le résultat. Villani

montre une majoration explicite :

$$\mathcal{W}_1(\mu, \nu) \leq Diam(\mathcal{X})\, \delta(\mu, \nu) \tag{3.5}$$

□

Chapitre 4

Applications en apprentissage statistique

Nous allons présenter deux applications différentes du transport optimal en apprentissage, qui ont émergé de la recherche ces dernières années, conséquence de l'apparition des nouvelles méthodes pour approximer une solution au problème du transport grâce notamment à l'algorithme du Sinkhorn.

4.1 Les réseaux antagonistes génératifs (GAN)

Contrairement à la plupart des modèles utilisés en apprentissage statistique qui sont souvent des modèles prédictifs, les réseaux antagonistes génératifs sont des modèles génératifs qui appartiennent à la catégorie de l'apprentissage non supervisé. À partir d'un grand nombre d'exemples issus d'une distribution inconnue, le GAN apprend à générer de nouveau exemples qui suivent la même distribution. Ces méthodes ont produits des résultats stupéfiants ces dernières années comme on peut le voir sur la figure 4.1.

FIGURE 4.1 – Exemples de visages générés par un réseau antagoniste génératif (StyleGAN2).

4.1.1 Fonctionnement des GAN

Les GAN s'appuient sur une structure particulière composée d'un générateur et d'un discriminateur. Le générateur a la tâche de générer des exemples indicernables des exemples provenant des données réelles, tandis que le discriminateur tente de différencier les exemples réels des exemples générés. L'entraînement est une compétition entre le générateur, qui essaie de tromper le discriminateur, et le discriminateur qui tente de ne pas être trompé.

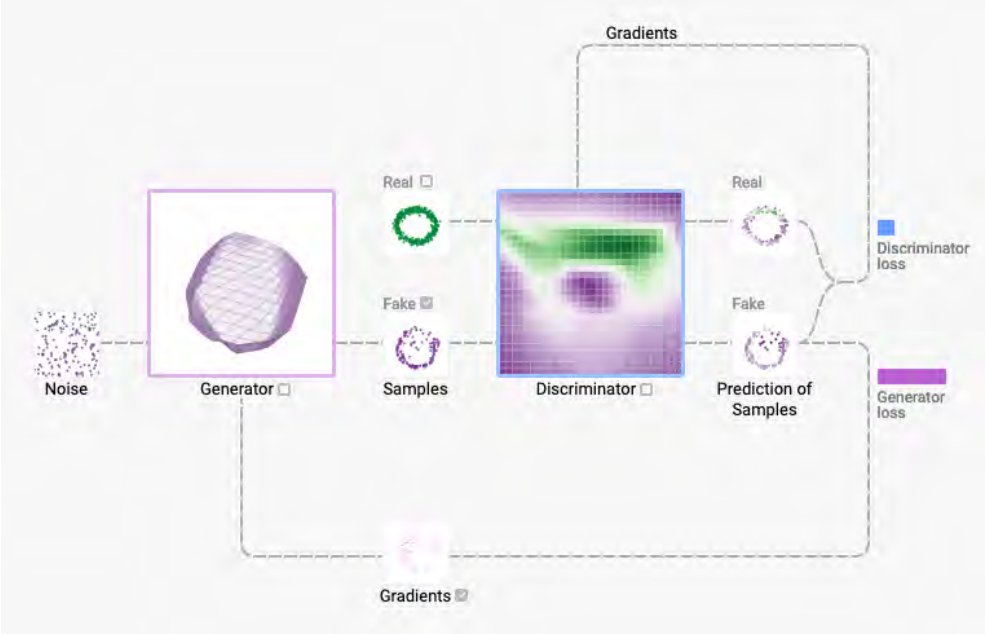


FIGURE 4.2 – Schéma du fonctionnement d'un GAN.

Rentrons un peu plus dans les détails. On note $\mu \in \mathrm{P}(\mathcal{X})$ la distribution des données réelles et que l'on cherche à estimer. On définit une variable aléatoire avec une distribution fixe ρ, dont on passe des réalisations par une fonction $g_\theta : \mathcal{Z} \to \mathcal{X}$ qui génère des exemples suivant une certaine distribution $\mu_\theta \in \mathrm{P}(\mathcal{X})$. La fonction g peut par exemple être un réseau neuronal paramétré par θ.

L'idée est de pouvoir mesurer la « proximité » entre les distributions μ et μ_θ. On cherche donc à définir une distance ou une divergence $l(\mu_\theta, \mu)$. Un des critères d'importance est l'impact de cette divergence ou cette distance sur la convergence des suites de distribution de probabilité.

On dit qu'une suite μ_n converge vers μ si, et seulement si, $l(\mu_n, \mu) \to 0$, et donc la notion de convergence associée dépend de l. Pour optimiser notre modèle

et donc optimiser le paramètre θ, il est souhaitable que l'application $\theta \to \mu_\theta$ soit continue, et donc que lorsque $\theta_n \to \theta, \mu_{\theta_n} \to \mu_\theta$. La distance ou divergence l doit donc préférablement induire une convergence la plus faible possible, qui garde du sens, afin qu'un grand nombre de suites convergent et qu'on puisse utiliser $l(\mu_\theta, \mu)$ comme une fonction de coût.

4.1.2 Apport du transport optimal

La réponse à la discussion commencée à la section 4.1.1 apportée par les GAN est d'utiliser une divergence telle que la divergence de Kullback-Leibler ou celle de Jensen-Shannon qui ont été définies au chapitre 3.

En 2017, Martin Arjovsky, Soumith Chintala et Léon Bottou dans [ACB17] apportent une amélioration conséquente aux GAN (désormais WGAN) en proposant d'utiliser à la place des divergences mentionnées plus haut une distance de Wasserstein. Comme discuté au chapitre 3, théorème 3.3.1, cette distance est plus naturelle que les divergences de Kullback-Leibler ou de Jensen-Shannon, elle induit une convergence plus faible (précisément la convergence en loi, ou convergence étroite) sur l'espace des distributions et permet également de beaucoup mieux comparer des distributions de supports différents.

Arjovsky, Chintala et Bottou ont également constaté des améliorations empiriques dans l'entraînement du GAN, telles qu'une meilleure stabilité, et une métrique qui a plus de sens et qui est mieux corrélée avec la qualité des exemples.

Nous montrons ci-après les résultats que nous avons obtenu en utilisant un WGAN sur l'ensemble de données MNIST, qui correspond à des chiffres manuscrits.

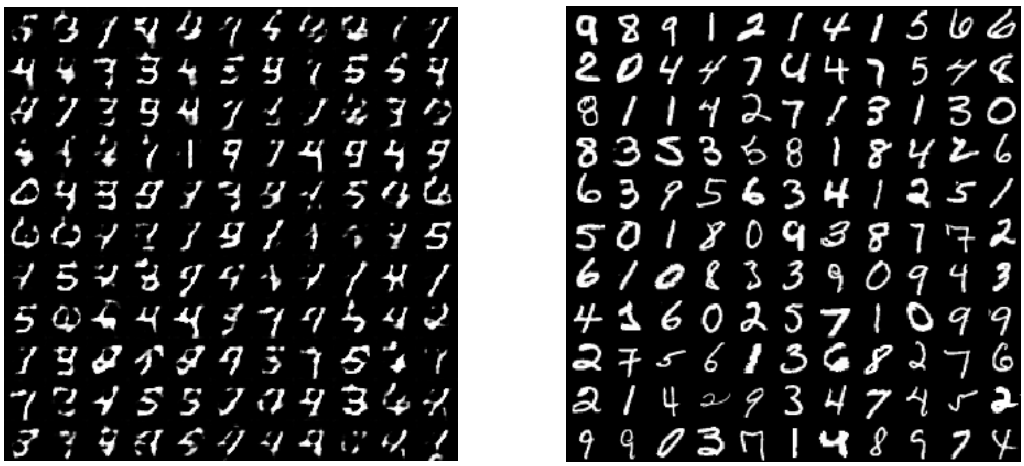


FIGURE 4.3 – Images générées par le WGAN au début de l'entraînement (à gauche) et à la fin (à droite).

4.2 Adaptation de domaine

4.2.1 Motivation du problème de l'adaptation de domaine

Il arrive souvent en apprentissage statistique que les données réelles ne soient pas distribuées tout à fait de la même manière que les données étiquetées utilisées pour l'entraînement du modèle. Par exemple en vision par ordinateur, cela peut être dû à une différence d'éclairage entre les données d'entraînement et réelles, ou alors une différence avec l'appareil d'acquisition ou encore la présence ou l'absence d'un fond. On peut facilement imaginer que ce genre de différences peut survenir dans d'autres contextes en apprentissage statistique. Le problème de l'adaptation de domaine est celui, à l'aide d'un modèle entraîné, d'obtenir les meilleures performances possibles sur les données réelles. Ces dernières années plusieurs équipes de recherches ont proposés des avancées encourageantes en utilisant le transport optimal.

FIGURE 4.4 – Exemple de situation motivant l'adaptation de domaine. À droite sont les données d'entraînement et à gauche les données réelles.

4.2.2 Formalisation du problème

Soit $\Omega \subset \mathbb{R}^d$ l'espace des entrées et $\mathcal{C}$ l'espace des étiquettes. On note $\mathrm{P}(\Omega)$ l'ensemble des mesures de probabilités sur Ω. Le paradigme d'apprentissage classique suppose l'existence d'un ensemble d'exemples étiquetés $\{(x_{train}, y_{train}) \in \Omega \times \mathcal{C}\}$ qui suivent la distribution jointe π, et d'un ensemble d'exemples $\{x_{test} \in \Omega\}$ dont

on ne connaît pas (et on souhaite déterminer) les étiquettes. Pour déterminer les étiquettes des x_{test} on utilise souvent une estimation empirique de π et on fait l'hypothèse que x_{train} et x_{test} suivent la même distribution.

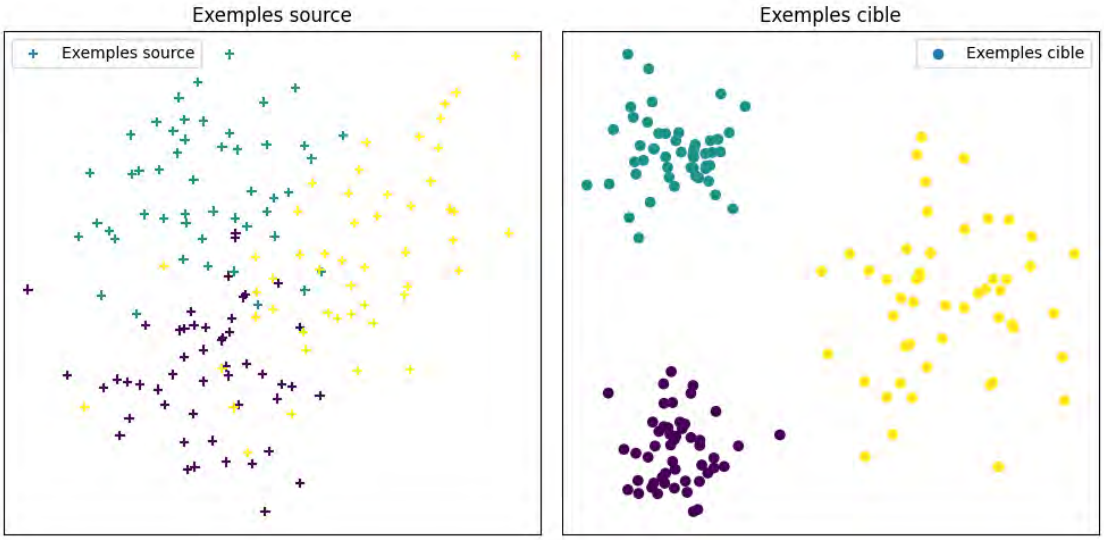
Le problème de l'adaptation de domaine ne fait pas cette dernière hypothèse et suppose au contraire l'existence de deux distributions jointes π_s et π_c relatives respectivement à un domaine source Ω_s et un domaine cible Ω_c. On note μ_s, μ_c les marginales respectives sur $\mathcal{X}$.

On suppose également que la différence de distribution des exemples dans le domaine source et cible provient d'une transformation $T : \Omega_s \to \Omega_c$ de l'espace des exemples. Finalement, on fait l'hypothèse que T préserve les distribution conditionnelles, c'est à dire que :

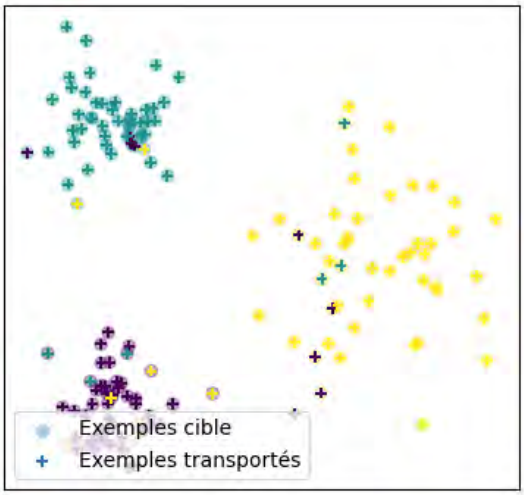
$$\pi_s(y \mid x^s) = \pi_c(y \mid T(x^s))$$

4.2.3 Apport du transport optimal

L'approche suggérée dans [CFTR15] est alors de chercher une application T qui transporte μ_s sur μ_c avec un coût minimal, le coût étant ici la distance euclidienne. Une fois ce plan de transport T calculé par les méthodes numériques exposées plus haut, on peut transporter les exemples sources dans l'espace cible Ω_c et ainsi entraîner un modèle directement dans l'espace cible. Nous présentons à la figure 4.5 les résultats d'expériences sur l'adaptation de domaine que nous avons menés sur un ensemble de données très simple. Les résultats très intéressants présentés dans [CFTR15] ont été suivis d'autres publications exploitant des idées similaires notamment [CFHR17], [DKF$^+$18] et l'utilisation des distances de Gromov-Wasserstein pour l'adaptation de domaines hétérogènes plus récemment ([VCF$^+$18]).



(a) Distribution des exemples dans l'espace source et dans l'espace cible.



(b) Distribution des exemples de l'espace source transportés dans l'espace cible.

FIGURE 4.5 – Exemple de l'utilisation du transport optimal pour l'adaptation de domaine.

Bibliographie

[ACB17] Arjovsky, Martin, Soumith Chintala et Léon Bottou: *Wasserstein GAN*, 2017.

[CFHR17] Courty, Nicolas, Rémi Flamary, Amaury Habrard et Alain Rakotomamonjy: *Joint Distribution Optimal Transportation for Domain Adaptation*, 2017.

[CFTR15] Courty, Nicolas, Rémi Flamary, Devis Tuia et Alain Rakotomamonjy: *Optimal Transport for Domain Adaptation*, 2015.

[Cut13] Cuturi, Marco: *Sinkhorn Distances : Lightspeed Computation of Optimal Transportation Distances*, 2013.

[DKF+18] Damodaran, Bharath Bhushan, Benjamin Kellenberger, Rémi Flamary, Devis Tuia et Nicolas Courty: *DeepJDOT : Deep Joint Distribution Optimal Transport for Unsupervised Domain Adaptation*, 2018.

[FL89] Franklin, Joel et Jens Lorenz: *On the scaling of multidimensional matrices*. Linear Algebra and Its Applications, pages 717–735, 1989.

[FZM+15] Frogner, Charlie, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo et Tomaso Poggio: *Learning with a Wasserstein Loss*, 2015.

[LBK17] Liu, Ming Yu, Thomas Breuel et Jan Kautz: *Unsupervised Image-to-Image Translation Networks*, 2017.

[PC18] Peyré, Gabriel et Marco Cuturi: *Computational Optimal Transport*, 2018.

[Tar97] Tarjan, Robert E.: *Dynamic Trees as Search Trees via Euler Tours, Applied to the Network Simplex Algorithm*. Math. Program., 78(2) :169–177, août 1997.

[TBGS17] Tolstikhin, Ilya, Olivier Bousquet, Sylvain Gelly et Bernhard Schoelkopf: *Wasserstein Auto-Encoders*, 2017.

[VCF+18] Vayer, Titouan, Laetita Chapel, Rémi Flamary, Romain Tavenard et Nicolas Courty: *Fused Gromov-Wasserstein distance for structured objects : theoretical foundations and mathematical properties*, 2018.

[Vil03] Villani, C.: *Topics in Optimal Transportation.* Graduate studies in mathematics. American Mathematical Society, 2003.

[Vil08] Villani, C.: *Optimal Transport : Old and New.* Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.

[Wen17] Weng, Lilian: *From GAN to WGAN.* lilianweng.github.io/lil-log, 2017.