BOULIN Alexis & KHALIFA Ottavio **ENSAE 3ème année**

Estimation paramétrique à l'aide de la distance de Wasserstein

Rapport de projet de Transport Optimal

Marco Cuturi **2020/2021**

1 Résumé de l'article

La distance de Wasserstein permet, à l'instar de la divergence de Kullback-Leibler, de mesurer la dissimilarité entre deux mesures de probabilité sur un espace donné. Elle est actuellement de plus en plus utilisée dans différents domaines de la statistique, avec par exemple en 2017 l'article [MA17] proposant de l'appliquer aux GAN. L'article présenté ici propose d'utiliser cette distance pour définir un estimateur dans le cadre des statistiques paramétriques. Après l'avoir défini, il propose un certain nombre de résultats théoriques, notamment de mesurabilité et de consistance, ainsi que plusieurs exemples d'utilisations pratiques. Nous allons ici résumer les principales idées et résultats théoriques présentés.

1.1 Notations

Dans toute la suite on se place dans un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. On note $\mathcal{P}(\mathcal{X})$ l'ensemble des mesures de probabilité sur un espace mesurable $\mathcal{X}$. On considèrera dans la suite des données à valeurs dans un sous-ensemble $\mathcal{Y}$ de $\mathbb{R}^d$.

Pour des données $y_1 \dots y_n$ dans $\mathcal{Y}$ distribuées selon une mesure de probabilté $\mu^* \in \mathcal{P}(\mathcal{Y}^n)$, on considère la distribution empirique :

$$\hat{\mu}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$$

On cherche à approcher cette distribution à l'aide d'un modèle paramétrique :

$$\mathcal{M}^{(n)} := \{\mu_\theta^{(n)}, \theta \in \mathcal{H}\} \subseteq \mathcal{P}(\mathcal{Y}^n)$$

où $\mathcal{H} \subseteq \mathbb{R}^{d_\theta}$ est l'espace des paramètres, que l'on munit d'une distance ρ.

On supposera souvent que pour tout θ dans $\mathcal{H}$, la suite $\hat{\mu}_{\theta,n}$ converge (dans un sens que l'on précisera) vers une distribution μ_θ, où :

$$\hat{\mu}_{\theta,n} = \frac{1}{n} \sum_{i=1}^{n} \delta_{z_i}$$

avec $z_i \sim \mu_\theta^{(n)}$. On considèrera alors $\mathcal{M} = \{\mu_\theta, \theta \in \mathcal{H}\}$ comme le modèle. Le modèle est dit bien spécifié s'il existe θ^* dans $\mathcal{H}$ tel que $\mu_* = \mu_{\theta^*}$.

Si ρ est une distance sur $\mathcal{Y}$, on pose $\mathcal{P}_p(\mathcal{Y})$ pour $p \geq 1$ comme l'ensemble des distributions μ de $\mathcal{P}(\mathcal{Y})$ telles que :

$$\exists y_0 \in \mathcal{Y}, \quad \int_{\mathcal{Y}} \rho(y, y_0)^p d\mu(y) < \infty$$

La p-distance de Wasserstein est alors une distance sur $\mathcal{P}_p(\mathcal{Y})$ définie par le problème de transport optimal suivant :

$$W_p(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{Y}\times\mathcal{Y}} \rho(x, y)^p d\gamma(x, y)$$

où $\Gamma(\mu, \nu)$ est l'ensemble des mesures de probabilité sur $\mathcal{Y} \times \mathcal{Y}$ de marginales μ et ν respectivement. L'un des avantages de la distance de Wasserstein comparée, par exemple, à la divergence de Kullback-Leibler, est qu'elle permet de comparer des distributions à supports disjoints.

Si f est une fonction à valeurs réelles possédant une borne inférieure, on pose $\varepsilon - \arg\min_x f = \{x, f(x) \leq \varepsilon + \inf_x f\}$.

Enfin, on note la convergence faible d'une suite (μ_n) de mesures vers μ par $\mu_n \Rightarrow \mu$.

1.2 Estimateurs MWE et MEWE

L'idée principale de l'article est d'utiliser la distance de Wasserstein pour minimiser la différence entre la distribution empirique $\hat{\mu}_n$ et la distribution modèle μ_θ. L'estimateur MWE (Minimum Wasserstein Estimation) est défini comme suit :

$$\hat{\theta}_n \in \underset{\theta \in \mathcal{H}}{\arg\min}\ W_p(\hat{\mu}_n, \mu_\theta)$$

En pratique, cette formule ne permet pas forcément de calculer l'estimateur. Pour les cas où, par exemple, on sait simuler des données selon μ_θ sans nécessairement connaître la formule de sa densité, on introduit l'estimateur MEWE (Minimum Expected Wasserstein Estimation) :

$$\hat{\theta}_{n,m} \in \underset{\theta \in \mathcal{H}}{\arg\min}\ \mathbb{E}_m[W_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})]$$

où l'espérance est obtenue à l'aide de l'échantillon simulé : $z_i \sim \mu_\theta^{(m)}$, donnant une distribution empirique : $\hat{\mu}_{\theta,m} = \sum_{i=1}^{m} \delta_{z_i}$.

1.3 Principaux résultats théoriques

1.3.1 Cas de l'estimateur MWE

On présente ici les premiers résultats d'existence, de mesurabilité et de consistance de l'estimateur MWE. On commence par faire trois hypothèses techniques.

Hypothèse 1. *$W_p(\hat{\mu}_n, \mu_*) \to 0$ $\mathbb{P}$-presque surement.*

Hypothèse 2. *L'application $\theta \mapsto \mu_\theta$ est continue, dans le sens où $\rho(\theta_n, \theta) \to 0$ implique que $\mu_{\theta_n} \Rightarrow \mu_\theta$.*

Hypothèse 3. *Il existe un $\varepsilon > 0$ telque l'ensemble $B_*(\varepsilon) = \{\theta \in \mathcal{H}, W_p(\mu_*, \mu_\theta) \leq \varepsilon_* + \varepsilon\}$ soit borné, où $\varepsilon_* = \inf_{\theta \in \mathcal{H}} W_p(\mu_*, \mu_\theta)$.*

On peut alors montrer les deux théorèmes suivants. Nous ne détaillerons pas intégralement leurs preuves, mais elles reposent toutes deux sur un résultat fondamental présenté dans [Vil08], que nous avons démontré en annexe (preuve du théorème A.1).

Théorème 1. *(Existence et consistence de l'estimateur MWE) Sous les hypothèses 1 à 3, il existe un ensemble $E \subseteq \Omega$, tel que $\mathbb{P}(E) = 1$ et tel que pour tout $\omega \in E$,*

$$\inf_{\theta \in \mathcal{H}} W_p(\hat{\mu}_n(\omega), \mu_\theta) \to \inf_{\theta \in \mathcal{H}} W_p(\mu_*, \mu_\theta)$$

De plus, il existe un entier $n(\omega)$ tel que pour tout $n \geq n(\omega)$, l'ensemble $\underset{\theta \in \mathcal{H}}{\arg\min} W_p(\mu_*, \mu_\theta)$ est non vide. Ces ensembles forment de plus une suite bornée, telle que :

$$\limsup_{n \to +\infty} \underset{\theta \in \mathcal{H}}{\arg\min}\ W_p(\hat{\mu}_n(\omega), \mu_\theta) \subseteq \underset{\theta \in \mathcal{H}}{\arg\min}\ W_p(\mu_*, \mu_\theta).$$

Théorème 2. *(Mesurabilité de l'estimateur MWE). On suppose que $\mathcal{H}$ est un borélien de $\mathbb{R}^{d_\theta}$ σ-compact. Sous l'hypothèse 2, pour tout $n \geq 1$ et tout $\varepsilon > 0$, il existe une fonction borélienne $\hat{\theta}_n : \Omega \to \mathcal{H}$ telle que :*

$$\hat{\theta}_n(\omega) \in \begin{cases} \underset{\theta \in \mathcal{H}}{\arg\min}\ W_p(\hat{\mu}_n(\omega), \mu_\theta) & \text{si cet ensemble est non-vide} \\ \varepsilon - \underset{\theta \in \mathcal{H}}{\arg\min}\ W_p(\hat{\mu}_n(\omega), \mu_\theta) & \text{sinon.} \end{cases}$$

1.3.2 Cas de l'estimateur MEWE

On peut trouver des résultats similaires pour MEWE, moyennant l'ajout de deux hypothèses.

Hypothèse 4. *Pour tout $m \geq 1$, si $\rho(\theta_n, \theta) \to 0$, alors $\mu_{\theta_n}^{(m)} \Rightarrow \mu_\theta^{(m)}$.*

Hypothèse 5. *Si $\rho(\theta_n, \theta) \to 0$, alors $\mathbb{E}_n[W_p(\mu_{\theta_n}, \hat{\mu}_{\theta_n,n}] \to 0$.*

Les résultats suivants sont analogues aux précédents. Pour simplifier, on considère m comme une fonction de n et on suppose que $m(n) \to +\infty$ quand $n \to +\infty$.

Théorème 3. *(Existence et consistence de l'estimateur MEWE) Sous les hypothèses* 1 *à* 5*, il existe un ensemble* $E \subseteq \Omega$*, avec* $\mathbb{P}(E) = 1$ *et tel que pour tout* $\omega \in E$ *:*

$$\inf_{\theta \in \mathcal{H}} \mathbb{E}_{m(n)} W_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) \to \inf_{\theta \in \mathcal{H}} W_p(\mu_*, \mu_\theta)$$

et il existe un entier $n(\omega)$ *tel que pour tout* $n \geq n(\omega)$*, l'ensemble* $\underset{\theta \in \mathcal{H}}{\arg\min}\ W_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}$ *est non vide. Ces ensembles forment de plus une suite bornée, telle que :*

$$\limsup_{n \to +\infty} \underset{\theta \in \mathcal{H}}{\arg\min}\ \mathbb{E}_{m(n)}[W_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})] \subseteq \underset{\theta \in \mathcal{H}}{\arg\min}\ W_p(\mu_*, \mu_\theta).$$

Théorème 4. *(Mesurabilité de l'estimateur MEWE) On suppose que* $\mathcal{H}$ *est un borélien de* $\mathbb{R}^{d_\theta}$ σ*-compact. Sous l'hypothèse* 4*, pour tous* $n \geq 1, m \geq 1$ *et* $\varepsilon > 0$*, il existe une fonction borélienne* $\hat{\theta}_{n,m} : \Omega \to \mathcal{H}$ *telle que :*

$$\hat{\theta}_{n,m}(\omega) \in \begin{cases} \underset{\theta \in \mathcal{H}}{\arg\min}\ \mathbb{E}_m[W_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m})] & \textit{si cet ensemble est non-vide} \\ \varepsilon - \underset{\theta \in \mathcal{H}}{\arg\min}\ \mathbb{E}_m[W_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m})] & \textit{sinon.} \end{cases}$$

De plus, sous une hypothèse supplémentaire, on peut montrer, à un nombre n de données fixé, que l'estimateur MEWE converge vers l'estimateur MWE lorsque celui-ci existe, quand m tend vers l'infini.

Hypothèse 6. *Il existe un* $\varepsilon > 0$*, tel que l'ensemble* $B_n(\varepsilon) = \{\theta \in \mathcal{H}, W_p(\hat{\mu}_n, \mu_\theta) \leq \varepsilon_n + \varepsilon\}$ *soit borné.*

Théorème 5. *(MEWE converge vers MWE) Sous les hypothèses* 2 *et* 4 *à* 6*, on a :*

$$\inf_{\theta \in \mathcal{H}} \mathbb{E}_m[W_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})) \underset{m \to +\infty}{\longrightarrow} \inf_{\theta \in \mathcal{H}} W_p(\hat{\mu}_n, \mu_\theta)$$

et il existe un entier $\hat{m}$ *tel que pour tout* $m \geq \hat{m}$*, l'ensemble* $\underset{\theta \in \mathcal{H}}{\arg\min}\ \mathbb{E}_m\ [W_p(\hat{\mu}_n, \hat{\mu}_{\theta),m}]$ *soit non vide. Ces ensembles forment de plus une suite bornée, telle que :*

$$\limsup_{m \to +\infty} \underset{\theta \in \mathcal{H}}{\arg\min}\ \mathbb{E}_m[W_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}] \subseteq \underset{\theta \in \mathcal{H}}{\arg\min}\ W_p(\hat{\mu}_n, \mu_\theta).$$

2 Implémentation

Nous avons implémenté les différents algorithmes calculant l'estimateur MEWE en python [1]. L'objet de la section II-A est d'estimer les paramètres lorsque les variables aléatoires sont à valeurs dans la ligne réelle. Dans ce cas particulier, la distance de Wasserstein entre les mesures empiriques $\mathbb{E}_m \mathcal{W}_1(\hat{\mu}_n, \hat{\mu}_{\theta,m})^1$ peut être exprimée selon :

$$\mathcal{W}_2^2(y_{1:n}, z_{1:m}) = \frac{1}{m} \sum_{i=1}^{n} \sum_{j=1}^{l} |y_{(i)} - z_{(l(i-1)+j)}|^2 \tag{1}$$

En ayant noté $m = nl$ pour $l \geq 1$. Ceci découle immédiatement de la représentation $\mathcal{W}_2^2(\hat{\mu}_n, \hat{\nu}_m) = \int_0^1 |F_{\mu,n}^{-1}(s) - F_{\nu,m}^{-1}(s)|^2$ démontrée en annexe.

La première simulation considérée est le cas de la section 4.2 du papier concernant la somme de variables aléatoires distribuées selon une loi log-Normale. Nous répliquons ensuite respectivement les sections 4.3 et 4.4 de l'article qui considèrent des modèles mal spécifiés. Le premier est l'estimation du paramètre de localisation et d'échelle d'une loi gamma. Pour celui-ci, on observe un échantillon provenant d'une loi $\Gamma(10, 5)$ et nous estimons ces paramètres en utilisant une mesure *a priori* gaussienne d'espérance μ et de variance σ^2. Pour le second, la mesure *a priori* est la même mais les données observées résultent d'une Cauchy de médiane 0 et un paramètre d'échelle valant 1.

Les estimations obtenues pour ces trois modèles résultent de l'algorithme 1. Sur le colab, les paramètres $M = 100$, $N = 20$, $M = 500$ et $n \in \{50, 100, 250, 500\}$ ont été choisies par des soucis de rapidité. En local, le même algorithme a été utilisé pour des valeurs plus grandes dont $M = 1000$, $N = 20$, $M = 10^4$, $n \in \{50, 100, 250, 500, 5000, 10000\}$ qui sont ceux utilisés dans l'article.

1. L'ensemble du travail informatique peut être trouvé sur le lien du colab.

Algorithm 1 MEWE estimator process on real line

procedure MEWE(M, N, m, n)
  System Initialization
  for $i = 0 : M$ **do**
    observe a realisation $y_{1:n}$ from $\mu_*^{(n)}$
    draws N sample from $\hat{\mu}_\theta^{(m)}$
    optimize $\frac{1}{N}\sum_{i=1}^{N} \mathcal{W}_1(\hat{\mu}_n, \hat{\mu}_{\theta,m}^{(i)})$ with respect to $\theta \in \mathcal{H}$
    store the value $\hat{\theta}_{n,m}$ the result of the optimization program in a output vector
  Return output vector

L'algorithme 1 n'étant valable que pour des variables aléatoires à valeur sur la ligne réelle, nous souhaitions prolonger l'analyse au delà. Ceci est l'objet de la section II-C pour l'implémentation et des sections III-E et III-F pour la simulation du colab. Pour ce faire, nous utilisons une version régularisée de la distance de Wasserstein :

$$\gamma^\zeta = \min_{\gamma \in \Gamma_{n,m}} \sum_{i=1}^{n}\sum_{j=1}^{m} \rho(y_i, z_j)\gamma_{ij} + \zeta \sum_{i=1}^{n}\sum_{j=1}^{m} \gamma_{ij} log \gamma_{ij}$$

incluant une pénalité sur l'entropie ζ. Le problème régularisé peut être résolu de façon itérative par le *sinkhorn algorithm* [Cut13].

Algorithm 2 MEWE estimator process beyond real line

procedure MEWE$(M, N, m, n, \zeta, niterations)$
  System Initialization
  for $i = 0 : M$ **do**
    observe a realisation $y_{1:n}$ from $\mu_*^{(n)}$
    draws N sample from $\hat{\mu}_\theta^{(m)}$
    solves regularized problem with sinkhorn algorithm to obtain $\hat{\mathcal{W}}_1(\hat{\mu}_n, \hat{\mu}_{\theta,m}^{(i)})$ for $i \in \{1, \dots, N\}$
    optimize $\frac{1}{N}\sum_{i=1}^{N} \hat{\mathcal{W}}_1(\hat{\mu}_n, \hat{\mu}_{\theta,m}^{(i)})$ with respect to $\theta \in \mathcal{H}$
    store the value $\hat{\theta}_{n,m}$ the result of the optimization program in a output vector
  Return output vector

Dans les applications, nous essayons d'estimer la moyenne d'un vecteur gaussien μ dont la matrice de covariance est supposée connue et donnée par la matrice identité. Dans la première simulation, nous proposons un vecteur gaussien de dimension 2 et l'algorithme tourne sur le colab suivant les paramètres : $M = 20$, $N = 20$, $m = 15$, $n =\in \{5, 10, 15\}$, $\zeta = 0.05$, $niterations = 100$ où $niterations$ désigne le nombre d'itération dans le *sinkhorn algorithm*. En local, le même algorithme est utilisé pour des paramètres plus grands (cf colab). Dans nos exercices, nous avons constaté qu'il s'agit principalement de l'étape 7 de l'algorithme qui nécessite du temps notamment pour des nuages avec beaucoup d'observation. La seconde simulation concerne un vecteur gaussien à valeur dans R^{10} permettant de vérifier si l'algorithme a bien été implémenté et pouvant se généraliser à des espaces plus grands.

Nous avons en dernière étape souhaité vérifier la robustesse de l'estimateur (section II-D pour l'implémentation et section III-G pour la simulation dans le colab). Nous avons estimer les paramètres d'un modèle gaussien contaminé bien spécifié en utilisant l'algorithme 1. Pour les deux types de contamination proposées, nous obtenons une mauvaise estimation de la variance et une augmentation de la variance de l'estimateur de la moyenne (mais qui est sans biais du fait de la symétrie de la distribution contaminée). Nous proposons une heuristique permettant de converser une bonne estimation de la moyenne tout en diminuant la variance et ceci dans dans un même échelle de temps d'éxécution de l'algorithme. Cette heuristique repose sur l'utilisation de la médiane des moyennes permettant de réduire l'impact des données contaminées en retravaillant l'échantillon.

Soit K un entier inférieur à n et soit $B_1, \dots, B_K$ une partition de $\{1, \dots, n\}$ en bloc de taille N/K (on suppose que K divise N sinon on retire des données). L'estimateur de la médiane des moyenne est donné par

$$\hat{\mu}_K^{MOM} := median(\hat{\mu}_{B_1}, \dots, \hat{\mu}_{B_K}) \tag{2}$$

où, pour tout $k \in \{1, \dots, K\}$, $\hat{\mu}_{B_k}$ est définie comme :

$$\hat{\mu}_{B_k} = \frac{K}{N} \sum_{i \in B_k} X_i$$

L'idée que nous proposons est la suivante. Nous souhaitons calculé l'estimateur :

$$\hat{\theta}_{K,m}^{MOM} = \underset{\theta \in \mathcal{H}}{argmin} \mathbb{E}_m \mathcal{W}_1(\hat{\mu}_K^{MOM}, \hat{\mu}_{\theta,m})$$

avec $\hat{\mu}_K^{MOM} = \frac{1}{K} \sum_{k=1}^{K} \delta_{\hat{\mu}_{10,k}^{MOM}}$ est la mesure empirique associé à l'échantillon constitué des K médianes des moyennes réalisées sur la partition de $\{1, \dots, n\}$ où, sur chaque bloc de taille n/K, on calcule une médiane des moyennes sur 10 blocs que l'on note $\hat{\mu}_{10,k}^{MOM}$ pour $k \in \{1, \dots, K\}$. Notons que le nombre d'observation doit être supérieur à $10K$ pour cette configuration. L'expression mathématique de $\hat{\mu}_{10,k}^{MOM}$ suit de celle de (2) :

$$\hat{\mu}_{10,k}^{MOM} := median(\hat{\mu}_{B_1^k}, \dots, \hat{\mu}_{B_{10}^k}) \tag{3}$$

avec, pour tout $s \in \{1, \dots, 10\}$:

$$\hat{\mu}_{B_s^k} = \frac{1}{card(B_s^k)} \sum_{i \in B_s^k} X_i$$

Algorithm 3 MEWE estimator process using median of means

procedure MEWE(M, N, m, n, K)
  System Initialization
  for $i = 0 : M$ **do**
    observe a realisation $y_{1:n}$ from $\mu_*^{(n)}$
    draws N sample from $\hat{\mu}_\theta^{(m)}$
    make a disjoint partition $B_1, \dots, B_K$ of $\{1, \dots, n\}$
    for $k = 1 : K$ **do**
      make a disjoint partition $(B_1^k, \dots, B_{10}^k)$ of B_k.
      compute $\hat{\mu}_{10,k}^{MOM} = median(\hat{\mu}_{B_1^k}, \dots, \hat{\mu}_{B_{10}^k})$
    optimize $\frac{1}{N} \sum_{i=1}^{N} \hat{\mathcal{W}}_1(\hat{\mu}_K^{MOM}, \hat{\mu}_{\theta,m}^{(i)})$ with respect to $\theta \in \mathcal{H}$
    store the value $\hat{\theta}_{n,m}$ the result of the optimization program in a output vector
  Return output vector

Nous faisons tourner l'algorithme 3 selon les paramètres $M = 100$, $N = 20$, $n \in \{50, 100, 250, 500\}$, $M = 20$ et $K = 500$. Cette procédure nous a permis de réduire la variance de l'estimateur de la moyenne pour notre modèle en présence de données contaminées. Ceci pour un temps d'éxécution proche à celui de l'algorithme 1. De plus, on constate qu'une augmentation du nombre de blocs K permet de réduire la variance de l'estimateur.

3 Conclusion

Suite à des travaux théoriques permettant de mieux comprendre le formalisme de l'article et l'intérêt de cet estimateur, nous avons reproduit trois exemples provenant de l'article et obtenu les mêmes résultats. Nous avons ensuite prolongé l'analyse dans des espaces de dimension supérieure et avons obtenu des conclusions similaires. Nous jugeons ensuite la robustesse de l'estimateur et nous proposons, dans une dernière approche, une heuristique permettant d'améliorer l'estimation en présence de données contaminées.

Annexe

Preuve de l'égalité (1)

Cette égalité peut être démontrée sans connaissances préalable de la théorie du transport optimal (un exemple de démonstration peut être obtenu dans le livre [BLM13] au chapitre 8). Je n'ai pas adopté ce point de vue, et j'introduis donc de nouveaux outils définissant le terme de monotonicité dans des ensembles. Toutes les définitions peuvent être retrouvées dans le chapitre 12 de [Roc70].

Definition 1. *(monotonicité) Un sous-ensemble $\Gamma \in \mathbb{R}^n \times \mathbb{R}^n$ est dit monotone si pour tout (x_1, y_1) et (x_2, y_2) dans Γ*

$$\langle y_2 - y_1, x_2 - x_1 \rangle \geq 0 \tag{4}$$

Remarque 1. *Un sous ensemble $\Gamma \in \mathbb{R} \times \mathbb{R}$ est monotone si et seulement si il est complètement ordonné par l'ordre partiel induit par le cône dans $\mathbb{R}^2_+$. En d'autres termes, si $\forall (x_1, y_1) \in \Gamma$ et $(x_2, y_2) \in \Gamma$, soit $x_1 \leq x_2$ et $y_1 \leq y_2$ ou $x_2 \leq x_1$ et $y_2 \leq y_1$.*

Definition 2. *(cyclical monotonicity) Un sous ensemble $\Gamma \in \mathbb{R}^n \times \mathbb{R}^n$ est monotone cyclique si pour tout choix de points $(x_1, y_1), \ldots, (x_m, y_m)$ (où $m \geq 1$) de Γ, nous avons :*

$$\sum_{i=1}^{m} \langle y_i, x_{i+1} - x_i \rangle \leq 0$$

avec la convention $x_{m+1} = x_1$.

Nous pouvons très vite tirer des définitions la remarque suivante.

Remarque 2. *Si un sous ensemble $\Gamma \in \mathbb{R} \times \mathbb{R}$ est monotone alors il est monotone cyclique. Il n'est pas immédiat, lorsque $n = 1$, qu'un sous ensemble monotone cyclique est monotone. Mais ce résultat est induit par le théorème qui suit.*

Théorème 6. *(Théorème de Rockafellar) Un sous ensemble non vide $\Gamma \subset \mathbb{R}^n \times \mathbb{R}^n$ est dit monotone cyclique si et seulement si il est inclu dans la sous différentielle d'une fonction convexe et semi-continue inférieurement φ sur $\mathbb{R}^n$.*

Remarque 3. *Si l'on se place sur la ligne réelle, tout ensemble monotone est inclue dans la sous différentielle d'une fonction semi-continue inférieurement. Cette caractérisation des ensembles monotones fait directement écho à la caractérisation des gradients des fonctions convexes avec les fonctions croissantes.*

Théorème 7. *(Critère d'optimalité de Knott-Smith) Soit μ, ν deux mesures de probabilité sur $\mathbb{R}^n$ ayant des moments d'ordres deux finis. On considère le problème de Kantorovich associé à la fonction de coût quadratique $c(x, y) = |x - y|^2$. Ainsi,*
$\pi \in \Pi(\mu, \nu)$ est optimal si et seulement si il existe une fonction convexe et semi continue inférieurement φ tel que

$$supp(\pi) \subset graph(\partial\varphi)$$

ou, de façon équivalente :

$$\forall (x, y) \quad \pi \; p.p, y \in \partial\varphi(x)$$

Théorème 8. *(théorème 2.18 de [Vil03]) Soit μ, ν deux mesures de probabilité sur $\mathbb{R}$ caractérisées respectivement par les fonctions de répartitions F et G. Soit π une mesure de probabilité définie sur $\mathbb{R}^2$ dont la distribution jointe est spécifiée par :*

$$H(x, y) = min\{F(x), G(y)\}$$

Alors π appartient à $\Pi(\mu, \nu)$ et est optimal pour la formulation de Kantorovich du problème de transport optimal pour la fonction de coût $c(x, y) = |x - y|^2$. De plus, la valeur du coût est donnée par :

$$\mathcal{W}_2^2(\mu, \nu) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^2 dt$$

Démonstration. On note par $F(x^-)$ la limite à gauche $\lim_{z\uparrow x} F(z)$. Cette limite existe par monotonicité de la fonction de répartition.

• Montrons que $\text{supp}(\pi) \subset \{(x,y) \in \mathbb{R}^2; F(x^-) \leq G(y) \text{ et } G(y^-) \leq F(x)\}$

Supposons, par l'absurde que $(x,y) \in \text{supp}(\pi)$ et $F(x^-) > G(y)$. Puisque G est continue à droite, que F et G sont toutes les deux croissantes, on peut prendre un voisinage $\mathcal{V}_x$ de x et un voisinage $\mathcal{V}_y$ de y aussi petit que possible tel que pour tout $(x',y') \in \mathcal{V}_x \times \mathcal{V}_y$ on ait $F(x') > G(y')$.

$$H(x',y') = \min\{F(x'), G(y')\} = G(y')$$

Ainsi, dans un rectangle centré en (x,y), la fonction H ne dépend que de la seconde variable y'. Ainsi, la mesure π ne charge pas ce rectangle et il suit que $(x,y) \notin \text{supp}(\pi)$.

• $\text{supp}(\pi)$ est un ensemble monotone.

Soit (x_1,x_2), $(y_1,y_2) \in \text{supp}(\pi)$. Supposons que $x_1 \geq x_2$, nous devons montrer que $y_1 \geq y_2$. Utilisons le premier point et le fait que la fonction F est croissante, nous avons les inégalités suivantes :

$$G(y_1) \geq F(x_1^-) \geq F(x_2) \geq G(y_2^-)$$

Si $G(y_1) \geq G(y_2^-)$ implique $y_2 \leq y_1$, alors le résultat et prouvé. Si ce n'est pas le cas, alors nécessairement $G(y_1) = F(x_1^-) = F(x_2) = G(y_2^-)$. Si $y_2 > y_1$ alors F est constante sur le segment $[x_2, x_1[$ et G sur le segment $[y_1, y_2[$. Montrons que ce cas n'est pas réalisable dans le sens où (x_1,y_1) et (x_2,y_2) ne peuvent pas appartenir au support de π. Prenons (x_2,y_2). Pour tout ϵ strictement positif, nous pouvons prendre un rectangle R dont les sommets sont $(x_2 \pm \epsilon, y_2 \pm \epsilon)$ dont la mesure pour π est zéro. En effet, en exprimant la mesure de ce rectangle par rapport à π, nous avons :

$$\pi(R) = H(x_2+\epsilon, y_2+\epsilon) + H(x_2-\epsilon, y_2-\epsilon) - H(x_2-\epsilon, y_2+\epsilon) - H(x_2+\epsilon, y_2-\epsilon)$$

En utilisant la définition de H, que $x_2 < x_1$, $y_2 > y_1$ et que F et G sont croissantes, nous obtenons le résultat. Ainsi, $y_1 > y_2$ est impossible et le deuxième point est démontré.

• Ainsi, le support de π est contenu dans un sous ensemble de $\mathbb{R}^2$ qui est monotone. Puisque nous somme situé sur la ligne réelle, les commentaires faits précédemment nous permettent d'affirmer que ce sous-ensemble est monotone cyclique. Par le théorème de Rockafellar, le support de π est ainsi contenu dans la sous différentielle d'une fonction convexe semi continue inférieurement. En appliquant le théorème de Knott-Smith, il suit que π est un plan de transport optimal. Nous souhaitons montré à présent que

$$\pi = (F^{-1} \times G^{-1})_{\#}\lambda \tag{5}$$

où λ désigne la mesure de lebesgue sur le segment $[0,1]$. Soit $R(x,y)$ un rectangle arbitraire, ainsi, (5) devient :

$$\pi[R(x,y)] = \lambda\{(F^{-1}(t), G^{-1}(t)) \in R(x,y)\}$$

Cette dernière quantité vaut :

$$\lambda(\{t \in \mathbb{R}, F^{-1}(t) \leq x\} \cap \{t \in \mathbb{R}, G^{-1}(t) \leq y\}) = H(x,y)$$

La classe :

$$\mathcal{E} = \{B \in \mathcal{B}(\mathbb{R}^2), \pi(B) = \lambda((F^{-1} \times G^{-1})^{-1}(B))\}$$

Contient donc la classe R de tous les rectangles fermés de $\mathbb{R}^2$. C'est par ailleurs une classe monotone (ou λ-système). Puisque la classe R des rectangles fermés de $\mathbb{R}^2$ est un π-système, le théorème de la classe monotone assure que $\mathcal{E}$ contient la tribu engendrée par R, c'est à dire que $\mathcal{E}$ contient $\mathcal{B}(\mathbb{R}^2)$. Nous obtenons alors l'identité (5).

• Une conséquence de (5) est que, pour chaque fonction mesurable positive u sur $\mathbb{R}^2$, nous avons :

$$\int_{\mathbb{R}^2} u(x,y) d\pi(x,y) = \int_0^1 u(F^{-1}(t), G^{-1}(t)) dt$$

Et, en particuler, en prenant $u(x,y) = |x-y|^2$, nous obtenons notre résultat. □

Preuve du théorème A.1

On propose ici une preuve du théorème A.1 de l'article, proposée par Villani dans [Vil08] (théorème 6.9). Ce théorème permet en particulier de montrer l'existence, la mesurabilité et la consistence des estimateurs MWE et MEWE, et est donc primordial pour la compréhension de l'article. On commence par quelques définitions.

Definition 3. *On dit qu'une suite de mesures (μ_n) de $\mathcal{P}_p(\mathcal{Y})$ converge faiblement dans $\mathcal{P}_p(\mathcal{Y})$ vers μ si les deux conditions suivantes sont vérifiées :*

(i) (μ_n) *converge faiblement vers* μ *au sens usuel, i.e pour toute fonction* f *continue et bornée sur* $\mathcal{Y}$, $\int_{\mathcal{Y}} f d\mu_n \underset{n\to+\infty}{\longrightarrow} \int_{\mathcal{Y}} f d\mu$

(ii) $\exists y_0 \in \mathcal{Y}, \int_{\mathcal{Y}} \rho(y, y_0)^p d\mu_n(y) \underset{n\to+\infty}{\longrightarrow} \int_{\mathcal{Y}} \rho(y, y_0)^p d\mu(y)$

On note alors $\mu_n \Rightarrow \mu$.

Definition 4. *Une fonction* $f : \mathcal{Y} \to \mathbb{R}$ *est dite semi-continue inférieurement en* x_0 *si* $\liminf_{x\to x_0} f(x) \geq f(x_0)$.

Definition 5. *Une famille de mesures* $(\mu_i)_{i\in I}$ *sur un espace topologique mesurable* X *est dite tendue si pour tout* ε *strictement positif, il existe un sous-ensemble compact de* X *, noté* K_ε *tel que* $\forall i \in I, \mu_i(X \backslash K_\varepsilon)$.

On rappelle un énoncé classique : le théorème de Prokhorov. Des détails de ce théorème peuvent être trouvés dans le premier chapitre du cours suivant [Lé21].

Théorème 9. *Si* $\mathcal{Y}$ *est un espace polonais, alors tout ensemble de mesures de probabilités sur* $\mathcal{Y}$ *est précompact pour la topologie faible si et seulement si il est tendu.*

On rappelle également un théorème fondamental attestant de la stabilité du transport optimal. Il se déduit du théorème de dualité de Kantorovitch.

Théorème 10. *Soient* $\mathcal{X}, \mathcal{Y}$ *deux espaces polonais, et* c *une fonction continue sur* $\mathcal{X} \times \mathcal{Y}$ *à valeurs réelles et minorée. Soit* $(c_k)_{k\in\mathbb{N}}$ *une suite de fonctions continues convergeant uniformément vers* c *sur* $\mathcal{X} \times \mathcal{Y}$, *et soient* $(\mu_k)_{k\in\mathbb{N}}$ *et* $(\nu_k)_{k\in\mathbb{N}}$ *deux suites de mesures de probabilités sur* $\mathcal{X}$ *et* $\mathcal{Y}$ *respectivement. On suppose que ces suites convergent respectivement vers des mesures de probabilités* μ *et* ν.

Pour tout entier k, *on considère* π_k *un plan de transport optimal entre* μ_k *et* ν_k, *pour la fonction de coût* c_k. *Si* $\forall k \in \mathbb{N}, \int_{\mathcal{X}\times\mathcal{Y}} c_k \, d\pi_k < +\infty$ *et* $\liminf_{k\to+\infty} \int_{\mathcal{X}\times\mathcal{Y}} c_k \, d\pi_k < +\infty$, *alors à extraction près la suite* (π_k) *converge faiblement vers une mesure* π *de marginales* μ *et* ν, *et* π *est un plan de transport optimal entre* μ *et* ν.

On considèrera dans la suite que $\mathcal{Y}$ est une partie de $\mathbb{R}^d$, et donc un espace polonais, que l'on munit d'une distance ρ qui en fait un espace complet. Voici l'énoncé du théorème que l'on souhaite démontrer :

Théorème 11. *La p-distance de Wasserstein métrise la convergence faible dans* $\mathcal{P}_p(\mathcal{Y})$ *i.e si* (μ_n) *est une suite de mesures de* $\mathcal{P}_p(\mathcal{Y})$, $\mu_n \Rightarrow \mu$ *si et seulement si* $W_p(\mu_n, \mu) \underset{n\to+\infty}{\longrightarrow} 0$.

Avant d'attaquer la démonstration, nous allons énoncer et démontrer quelques lemmes.

Lemme 1. *Si* $p \geq 1$ *et si* $(\mu_k)_{k\in\mathbb{N}}$ *est une suite de Cauchy dans l'espace métrique* $(\mathcal{P}_p(\mathcal{Y}), W_p)$, *alors* $(\mu_k)_{k\in\mathbb{N}}$ *est tendue.*

Démonstration. $(\mu_k)_{k\in\mathbb{N}}$ est de Cauchy dans $(\mathcal{P}_p(\mathcal{Y}), W_p)$, i.e $W_p(\mu_k, \mu_l) \underset{k,l\to+\infty}{\longrightarrow} 0$.

$W_1 \leq W_p$ donc $(\mu_k)_{k\in\mathbb{N}}$ est aussi de Cauchy pour la distance W_1. On peut donc écrire :

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall k \geq N, W_1(\mu_N, \mu_k) \leq \varepsilon^2.$$

Donc $\forall k \in \mathbb{N}, \exists j \in [\![1, N]\!], W_1(\mu_j, \mu_k) \leq \varepsilon^2$ (par la ligne précédente si $k \geq N$, sinon prendre $j = k$).

$\mathcal{Y}$ est polonais, donc d'après le théorème de Prokhorov, l'ensemble fini $(\mu_1 \dots \mu_N)$ est tendu. Donc il existe un compact K de $\mathcal{Y}$ tel que $\forall j \in [\![1, N]\!], \mu_j(\mathcal{Y}\backslash K) \leq \varepsilon$.

Par compacité de K on a :

$$\exists m \in \mathbb{N}, \exists (x_1 \dots x_m) \in \mathcal{Y}^m, K \subseteq \bigcup_{i=1}^{m} B(x_i, \varepsilon).$$

En posant $U = \bigcup_{i=1}^{m} B(x_i, \varepsilon)$, on considère $U_\varepsilon = \{x \in \mathcal{Y}, \rho(x, U) \leq \varepsilon\}$. On a $U_\varepsilon \subseteq \bigcup_{i=1}^{m} B(x_i, 2\varepsilon)$.

On considère la fonction ϕ définir sur $\mathcal{Y}$ par : $\phi(x) = \left(1 - \frac{\rho(x,U)}{\varepsilon}\right)_+$. On remarque immédiatement que $\mathbb{1}_U \leq \phi \leq \mathbb{1}_{U_\varepsilon}$.

On rappelle la classique formule de dualité pour la distance de Kantorovitch-Rubinstein :

$$W_1(\mu, \nu) = \sup_{\mathrm{Lip}(\psi) \leq 1} \left(\int_{\mathcal{Y}} \psi \, d\mu - \int_{\mathcal{Y}} \psi \, d\nu \right). \tag{6}$$

On a donc : si $j \leq N$ et $k \in \mathbb{N}$ quelconque :

$$\begin{aligned}
\mu_k(U_\varepsilon) &= \int_{\mathcal{Y}} \mathbb{1}_{U_\varepsilon} d\mu_k \\
&\geq \int_{\mathcal{Y}} \phi \, d\mu_k = \int_{\mathcal{Y}} \phi \, d\mu_j + \left(\int_{\mathcal{Y}} \phi \, d\mu_k - \int_{\mathcal{Y}} \phi \, d\mu_j \right) \\
&\geq \int_{\mathcal{Y}} \phi \, d\mu_j - \frac{W_1(\mu_k, \mu_j)}{\varepsilon} \qquad \text{(d'après (3) et car } \phi \text{ est } \frac{1}{\varepsilon}\text{-lipschitzienne.)} \\
&\geq \int_{\mathcal{Y}} \mathbb{1}_U \, d\mu_j - \frac{W_1(\mu_k, \mu_j)}{\varepsilon} = \mu_j(U) - \frac{W_1(\mu_k, \mu_j)}{\varepsilon}.
\end{aligned}$$

On a : $K \subseteq U$ donc $\mu_j(U) \geq \mu_j(K) \geq 1 - \varepsilon$, et $W_1(\mu_j, \mu_k) \leq \varepsilon^2$. En reportant dans l'inégalité ci-dessus on obtient $\mu_k(U_\varepsilon) \geq 1 - 2\varepsilon$. On a donc :

$$\forall \varepsilon > 0, \exists (x_1 \dots x_m) \in \mathcal{Y}^m, \forall k \in \mathbb{N}, \mu_k \left(\bigcup_{i=1}^{m} \overline{B(x_i, 2\varepsilon)} \right) \geq 1 - 2\varepsilon.$$

$Z = \bigcup_{i=1}^{m} \overline{B(x_i, 2\varepsilon)}$ est bien un compact de $\mathcal{Y}$, ce qui conclut la preuve. □

Lemme 2. *Soient $\mathcal{X}, \mathcal{Y}$ deux espaces polonais et $c : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R} \cup \{+\infty\}$ une fonction de coût semi-continue supérieurement.*

Si $h : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R} \cup \{-\infty\}$ est une fonction semi-continue supérieurement telle que $c \geq h$ et $(\pi_k)_{k \in \mathbb{N}}$ est une suite de mesures de probabilité sur $\mathcal{X} \times \mathcal{Y}$ convergeant faiblement vers π, avec $h \in L^1(\pi_k)$ pour tout k et $h \in L^1(\pi)$, et si :

$$\int_{\mathcal{X} \times \mathcal{Y}} h \, d\pi_k \underset{k \to +\infty}{\longrightarrow} \int_{\mathcal{X} \times \mathcal{Y}} h \, d\pi$$

Alors :

$$\int_{\mathcal{X} \times \mathcal{Y}} c \, d\pi \leq \liminf_{k \to +\infty} \int_{\mathcal{X} \times \mathcal{Y}} c \, d\pi_k$$

Démonstration. Quitte à remplacer c par $c - h$, on peut supposer que $c \geq 0$. On peut écrire c comme limite simple d'une suite croissante de fonctions continues à valeurs réelles $(c_l)_{l \in \mathbb{N}}$. Le théorème de convergence monotone et le lemme de Fatou donnent alors :

$$\begin{aligned}
\int_{\mathcal{X} \times \mathcal{Y}} c \, d\pi &= \lim_{l \to +\infty} \int_{\mathcal{X} \times \mathcal{Y}} c_l \, d\pi \\
&= \lim_{l \to +\infty} \lim_{k \to +\infty} \int_{\mathcal{X} \times \mathcal{Y}} c_l \, d\pi_k \\
&\leq \liminf_{k \to +\infty} \int_{\mathcal{X} \times \mathcal{Y}} c \, d\pi_k.
\end{aligned}$$

□

Lemme 3. *Soient $\mathcal{X}, \mathcal{Y}$ deux espaces polonais, et $P \subseteq \mathcal{P}(\mathcal{X}), Q \subseteq \mathcal{P}(\mathcal{Y})$ deux sous-ensembles tendus. Alors l'ensemble $\pi(P,Q)$ des plans de transférence dont les marginales sont respectivement dans P et Q est tendu dans $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$.*

Démonstration. Soit $\mu \in P, \nu \in Q, \pi \in \pi(\mu,\nu)$. Par hypothèse, pour tout $\varepsilon > 0$, il existe un compact K_ε de $\mathcal{X}$, indépendant du choix de μ, et un compact L_ε de $\mathcal{Y}$ indépendant du choix de ν, tels que $\mu(\mathcal{X}\backslash K_\varepsilon) \leq \varepsilon$ et $\nu(\mathcal{Y}\backslash L_\varepsilon) \leq \varepsilon$.

Si (X,Y) est une variables aléatoire sur $\mathcal{X}, \mathcal{Y}$ de marginales μ, ν, alors :

$$\mathbb{P}((X,Y) \notin (K_\varepsilon, L_\varepsilon)) \leq \mathbb{P}(X \notin K_\varepsilon) + \mathbb{P}(Y \notin L_\varepsilon) \leq 2\varepsilon.$$

$(K_\varepsilon, L_\varepsilon)$ étant compact dans $\mathcal{X} \times \mathcal{Y}$, le résultat suit. □

Lemme 4. *Une suite de mesures (μ_n) de $\mathcal{P}_p(\mathcal{Y})$ converge faiblement vers μ dans $\mathcal{P}_p(\mathcal{Y})$ si et seulement si elle converge faiblement vers μ et :*

$$\exists x_0 \in \mathcal{Y}, \limsup_{k\to+\infty} \int_{\mathcal{Y}} \rho(x,x_0)^p \ d\mu_k(x) \leq \int_{\mathcal{Y}} \rho(x,x_0)^p \ d\mu(x)$$

Démonstration. Le sens direct est évident. Pour le sens indirect, on applique simplement le lemme 2 :

$$\limsup_{k\to+\infty} \int_{\mathcal{Y}} \rho(x,x_0)^p \ d\mu_k(x) \leq \int_{\mathcal{Y}} \rho(x,x_0)^p \ d\mu(x) \leq \liminf_{k\to+\infty} \int_{\mathcal{Y}} \rho(x,x_0)^p \ d\mu_k(x)$$

Les limites inférieure et supérieure étant égales, on trouve bien $\int_{\mathcal{Y}} \rho(x,x_0)^p \ d\mu_k(x) \underset{k\to+\infty}{\longrightarrow} \int_{\mathcal{Y}} \rho(x,x_0)^p \ d\mu(x)$.
□

On peut à présent passer à la preuve proprement dite du théorème.

Preuve du théorème

On commence par le sens indirect. Soit (μ_k) une suite d'éléments de $\mathcal{P}_p(\mathcal{Y})$ telle que $W_p(\mu_k,\mu) \underset{k\to+\infty}{\longrightarrow} 0$. Montrons que $\mu_k \Rightarrow \mu$.

D'après le lemme 1, (μ_k) étant une suite de Cauchy dans $(\mathcal{P}_p(\mathcal{Y}), W_p)$, elle est tendue. D'après le théorème de Prokhorov, il existe donc une sous-suite $(\mu_{\phi(k)})$ qui converge faiblement vers $\tilde{\mu} \in \mathcal{P}_p(\mathcal{Y})$.

D'après le lemme 2, on a : $W_p(\tilde{\mu},\mu) \leq \liminf_{k\to+\infty} W_p(\mu_{\phi(k)},\mu) = 0$. Donc $\tilde{\mu} = \mu$ et (μ_k) converge faiblement vers μ.

Si b est un réel positif fixé, alors $\forall \varepsilon > 0, \forall P \geq 1, (a+b)^p - (1+\varepsilon)a^p \underset{a\to+\infty}{\longrightarrow} -\infty$. De plus, $\frac{(a+b)^p - (1+\varepsilon)a^p}{b^p} \underset{b\to+\infty}{\longrightarrow} 1$, et $\frac{(a+b)^p - (1+\varepsilon)a^p}{b^p} \underset{b\to 0}{\longrightarrow} -\infty$.

Donc l'ensemble $\left\{ \frac{(a+b)^p - (1+\varepsilon)a^p}{b^p}, a \geq 0, b > 0 \right\}$ est majoré, et par conséquent :

$$\forall \varepsilon > 0, \exists C_\varepsilon > 0, \forall a,b \geq 0, (a+b)^p \leq (1+\varepsilon)a^p + C_\varepsilon b^p$$

On a donc :

$$\forall x_0, x, y \in \mathcal{Y}, \rho(x,x_0)^p \leq (\rho(x_0,y) + \rho(y,x))^p \leq (1+\varepsilon)\rho(x_0,y)^p + C_\varepsilon \rho(x,y)^p.$$

On considère pour tout entier k un plan de transport optimal π_k entre μ_k et μ. En intégrant l'inégalité ci-dessus selon π_k et en utilisant les propriétés de marginalité, il vient :

$$\begin{aligned}
\int_{\mathcal{Y}\times\mathcal{Y}} \rho(x_0,x)^p \ d\pi_k(x,y) &= \int_{\mathcal{Y}} \rho(x_0,x)^p \ d\mu_k(x) \\
&\leq (1+\varepsilon) \int_{\mathcal{Y}} \rho(x_0,y)^p \ d\mu(y) + C_\varepsilon \int_{\mathcal{Y}} \rho(x,y)^p \ d\pi_k(x,y) \\
&= (1+\varepsilon) \int_{\mathcal{Y}} \rho(x_0,y)^p \ d\mu(y) + C_\varepsilon W_p(\mu_k,\mu)^p.
\end{aligned}$$

Or par hypothèse $W_p(\mu_k,\mu)^p \underset{k\to+\infty}{\longrightarrow} 0$. En passant à la limite on obtient donc :

$$\limsup_{k\to+\infty} \int_{\mathcal{Y}} \rho(x_0,x)^p \, d\mu_k(x) \leq (1+\varepsilon) \int_{\mathcal{Y}} \rho(x_0,x)^p \, d\mu(x)$$

D'après le lemme 4, on a bien $\mu_k \Rightarrow \mu$.

Pour le sens indirect, on suppose que $\mu_k \Rightarrow \mu$. Comme précédemment, on considère pour tout entier k un plan de transport optimal π_k entre μ_k et μ.

D'après le théorème de Prokhorov, la suite (μ_k) est tendue. L'ensemble $\{\mu\}$ l'est également. D'après le lemme 3, la suite (π_k) est donc tendue dans $\mathcal{P}(\mathcal{Y}\times\mathcal{Y})$, et possède donc une sous-suite qui converge. Donc d'après le théorème de stabilité du transport optimal, à extraction près, la suite (π_k) converge faiblement vers une mesure π dans $\mathcal{P}(\mathcal{Y}\times\mathcal{Y})$, qui est un plan de transport optimal entre μ et μ. π est donc nécessairement le plan de transférence trivial $(Id,Id)_{\#\mu}$.

Soit $x_0 \in \mathcal{Y}$ et $R>0$. Si $x,y \in \mathcal{Y}$ sont tels que $\rho(x,y) > R$, alors : $\max(\rho(x,x_0),\rho(x_0,y)) \geq \dfrac{\rho(x,y)}{2}$ et $\max(\rho(x,x_0),\rho(x_0,y)) \geq \dfrac{R}{2}$ Dit autrement cela donne :

$$\mathbb{1}_{\rho(x,y)\geq R} \leq \mathbb{1}_{\rho(x,x_0)\geq \frac{R}{2}} \mathbb{1}_{\rho(x,x_0)\geq \frac{\rho(x,y)}{2}} + \mathbb{1}_{\rho(x_0,y)\geq \frac{R}{2}} \mathbb{1}_{\rho(x_0,y)\geq \frac{\rho(x,y)}{2}}.$$

En multipliant par $[\rho(x,y)^p - R^p]_+$, on obtient :

$$\begin{aligned}
[\rho(x,y)^p - R^p]_+ &\leq \rho(x,y)^p \mathbb{1}_{\rho(x,x_0)\geq \frac{R}{2}} \mathbb{1}_{\rho(x,x_0)\geq \frac{\rho(x,y)}{2}} + \rho(x,y)^p \mathbb{1}_{\rho(x_0,y)\geq \frac{R}{2}} \mathbb{1}_{\rho(x_0,y)\geq \frac{\rho(x,y)}{2}} \\
&\leq 2^p \rho(x,x_0)^p \mathbb{1}_{\rho(x,x_0)\geq \frac{R}{2}} + 2^p \rho(y,x_0)^p \mathbb{1}_{\rho(y,x_0)\geq \frac{R}{2}}.
\end{aligned}$$

On a alors :

$$\begin{aligned}
W_p(\mu_k,\mu)^p &= \int_{\mathcal{Y}} \rho(x,y)^p d\pi_k(x,y) \\
&= \int_{\mathcal{Y}} \min(\rho(x,y),R)^p d\pi_k(x,y) + \int_{\mathcal{Y}} [\rho(x,y)^p - R^p]_+ d\pi_k(x,y). \\
&\leq \int_{\mathcal{Y}} \min(\rho(x,y),R)^p d\pi_k(x,y) + 2^p \int_{\rho(x,x_0)\geq\frac{R}{2}} \rho(x,x_0)^p d\pi_k(x,y) + 2^p \int_{\rho(y,x_0)\geq\frac{R}{2}} \rho(y,x_0)^p d\pi_k(x,y) \\
&= \int_{\mathcal{Y}} \min(\rho(x,y),R)^p d\pi_k(x,y) + 2^p \int_{\rho(x,x_0)\geq\frac{R}{2}} \rho(x,x_0)^p d\mu_k(x) + 2^p \int_{\rho(y,x_0)\geq\frac{R}{2}} \rho(y,x_0)^p d\mu(y)
\end{aligned}$$

(π_k) converge faiblement vers π, donc le premier terme est positif et majoré par $W_p(\mu,\mu) = 0$ quand $k \to +\infty$, tend donc vers 0. Le côté gauche de l'inégalité ne dépendant pas de R, on peut alors écrire :

$$\begin{aligned}
\limsup_{k\to+\infty} W_p(\mu_k,\mu)^p &\leq \lim_{R\to+\infty} \limsup_{k\to+\infty} \left[\int_{d(x,x_0)\geq\frac{R}{2}} d(x,x_0)^p d\mu_k(x) + \int_{d(x_0,y)\geq\frac{R}{2}} \int_{d(x_0,y)} d(x_0,y)^p d\mu(y) \right] \\
&= 0 \text{ (en effet les bornes d'intégration tendent vers l'infini)}
\end{aligned}$$

Ceci achève la preuve du théorème.

Références

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities. A nonasymptotic theory of independence*. Oxford University Press, 2013.

[Cut13] Marco Cuturi. Sinkhorn distances : Lightspeed computation of optimal transportation distances, 2013.

[Lé21] Thierry Lévy. *Convergences de mesures, Grandes déviations, Percolations*. Sorbonne Université, 2021.

[MA17] Léon Bottou Martin Arjovsky, Soumith Chintala. Wasserstein gan. 01 2017.

[Roc70] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.

[Vil03] C. Villani. Topics in optimal transportation theory. 58, 01 2003.

[Vil08] Cédric Villani. *Optimal transport, old and new*. 2008.