Approximation du transport optimal entre distributions empiriques par flux de normalisation

Florentin COEURDOUX[1], Nicolas DOBIGEON[1] et Pierre CHAINAIS[2]

[1]University of Toulouse, IRIT/INP-ENSEEIHT, 31071 Toulouse Cedex 7, France

[2]Ecole Centrale de Lille, CRIStAL, 59651 Villeneuve d'Ascq, France

Florentin.Coeurdoux@irit.fr, Nicolas.Dobigeon@irit.fr
pierre.chainais@centralelille.fr

Résumé – Les flux de normalisation sont des outils génériques et puissants pour l'élaboration de modèles probabilistes et l'estimation de densité. Dans cet article, nous montrons que cette classe de modèles permet également d'approcher la solution d'un problème de transport optimal entre distributions empiriques quelconques. Précisément, la fonction de transport optimal est approchée par un réseau inversible dont l'entraînement repose sur une relaxation de la formulation de Monge. Cette approche a notamment l'avantage de permettre une discrétisation de cette fonction de transport en une composition de fonctions associées à chacune des couches du réseau, permettant d'obtenir les transports intermédiaires entre deux mesures.

Abstract – Normalization flows are generic and powerful tools for probabilistic modeling and density estimation. In this paper, we show that this class of models can also be used to approximate the solution of an optimal transport problem between any empirical distributions. Specifically, the optimal transport plan is approximated by an invertible network whose training is based on the relaxation of the Monge formulation. This approach has the advantage of allowing a discretization of this transport plan into a composition of functions associated with each layer of the network, providing intermediate transports between two measures.

1 Introduction

Le problème de transport optimal (TO) a été énoncé par le géomètre français Gaspard Monge. Dans son article séminal publié en 1781 [1], il formule la question suivante : comment déplacer un tas de terre vers un emplacement cible avec le moins d'effort ou de coût possible ? L'objectif était de trouver la meilleure façon de minimiser ce coût par une fonction de transport, sans devoir énumérer tous les appariements possibles entre les points de départ et les points d'arrivée. Récemment, le TO a trouvé des applications nombreuses en analyse ou traitement de données dans des domaines aussi variés que le traitement d'image ou l'apprentissage automatique [2].

Par ailleurs, les flux de normalisation (NF pour *normalizing flows*) ont récemment suscité beaucoup d'intérêt au sein de la communauté de l'apprentissage automatique, motivé notamment par leur capacité à modéliser facilement des données de grande dimension. Ces réseaux profonds sont caractérisés par un opérateur inversible qui associe à n'importe quelle distribution de données d'entrée une distribution cible qui est généralement choisie gaussienne centrée réduite. Parmi ses applications, on retrouve notamment la génération d'images avec RealNVP [3] ou Glow [4].

Motivés par les similitudes entre l'entraînement des NF et le problème de TO, nous proposons une architecture neuronale ainsi qu'une stratégie d'entraînement correspondante qui permet d'approcher la fonction de transport entre deux ensembles de points de distributions quelconques. La méthode proposée repose sur une relaxation de la formulation de Monge du TO. Cette fonction de coût, complétée par une régularisation de Sobolev adaptée à la structure en couches du réseau, est minimisée pour ajuster les poids du NF. Des simulations numériques montrent que cette régularisation permet d'obtenir une trajectoire plus régulière dont la discrétisation fournit implicitement des transports intermédiaires.

La Section 2 rappelle la formulation de Monge du TO et propose une relaxation dans le cas d'un transport entre deux distributions empiriques. La Section 3 présente le cadre générique des NF et décrit une instance particulière pour résoudre le problème de TO. Enfin, la Section 4 présente quelques résultats d'expériences illustrant les performances de la méthode proposée. La Section 5 conclut cette article.

2 Formulation de Monge du transport optimal et relaxation proposée

Soient μ et ν deux mesures de probabilités avec un second moment fini. Des mesures plus générales, par exemple sur $\mathcal{X} = \mathbb{R}^d$ (où $d \in \mathbb{N}^*$ est la dimension), peuvent avoir une densité $\mathrm{d}\mu(x) = p_X(x)dx$ par rapport à la mesure de Lebesgue, sou-

Ce travail a été soutenu par Artificial Natural Intelligence Toulouse Institute (ANITI, ANR-19-PI3A-0004), la Chaire IA Sherlock (ANR-20-CHIA-0031-01), le programme national d'investissement d'avenir ULNE (ANR-16-IDEX-0004) et la Région Hauts-de-France.

vent notée $p_X = \frac{\mathrm{d}\mu}{\mathrm{d}x}$, ce qui signifie que

$$\forall h \in \mathcal{C}\left(\mathbb{R}^d\right), \quad \int_{\mathbb{R}^d} h(x)\mathrm{d}\mu(x) = \int_{\mathbb{R}^d} h(x)p_X(x)\mathrm{d}x$$

où $\mathcal{C}(\cdot)$ est la classe des fonctions continues. Dans la suite de cet article, nous pourrons utiliser de manière interchangeable $\mathrm{d}\mu(x)$ et $p_X(x)\mathrm{d}x$.

2.1 Transport optimal selon Monge

Soient $\mathcal{X}$ et $\mathcal{Y}$ deux espaces métriques séparables. On rappelle qu'à toute application mesurable $T : \mathcal{X} \to \mathcal{Y}$ on peut associer son extension $T_\sharp$ qui déplace une mesure de probabilité sur $\mathcal{X}$ vers une nouvelle mesure de probabilité sur $\mathcal{Y}$. À une mesure μ sur $\mathcal{X}$ on associe la mesure image $\nu = T_\sharp\mu$ sur $\mathcal{Y}$ telle que

$$\forall h \in \mathcal{C}(\mathcal{Y}), \quad \int_{\mathcal{Y}} h(y)\mathrm{d}\nu(y) = \int_{\mathcal{X}} h(T(x))\mathrm{d}\mu(x).$$

Intuitivement, l'application $T : \mathcal{X} \to \mathcal{Y}$ peut être interprétée comme une fonction déplaçant un point unique d'un espace mesurable à un autre [2]. L'opérateur $T_\sharp$ pousse chaque masse élémentaire d'une mesure μ sur $\mathcal{X}$ en appliquant la fonction T pour obtenir alors une masse élémentaire dans $\mathcal{Y}$. Nous rappelons alors l'énoncé du problème de TO par Monge dans un cadre général. Pour une fonction de coût $c : \mathcal{X} \times \mathcal{Y} \to [0, +\infty]$, l'application mesurable $T : \mathcal{X} \to \mathcal{Y}$ est appelée fonction de transport optimal d'une mesure μ vers la mesure image $T_\#\mu = \nu$ si elle atteint l'infimum

$$\inf_T \left\{ \int_{\mathcal{X}} c(x, T(x))\mathrm{d}\mu(x) : T_\sharp\mu = \nu \right\}. \tag{1}$$

2.2 Relaxation de la formulation de Monge

Le transport optimal est un problème variationnel, c'est-à-dire nécessitant la minimisation d'un critère intégral dans une classe de fonctions admissibles. Étant données deux mesures de probabilité μ et ν, l'existence et l'unicité d'un opérateur T appartement à la classe des fonctions bijectives, continues et dérivables tel que $T_\sharp\mu = \nu$ ne sont pas garanties. La difficulté réside alors dans la classe définissant ces fonctions admissibles. En effet, même lorsque μ et ν sont des densités régulières sur des sous-ensembles eux-aussi réguliers de $\mathbb{R}^d$, la recherche d'une fonction de transport telle que $T_\sharp\mu = \nu$ rend le problème (1) difficile dans un cas général. Pour contourner la difficulté de la résolution de cette équation sur $T_\sharp$, nous proposons de reformuler l'énoncé de Monge en relaxant l'égalité sur l'opérateur définissant la mesure image.

Plus précisément, l'égalité entre la mesure image $T_\sharp\mu$ et la mesure cible ν est remplacée par la minimisation de leur distance $d(T_\sharp\mu, \nu)$. Le choix de la distance statistique $d(\cdot, \cdot)$ est crucial car c'est elle qui détermine la qualité de l'approximation de la mesure image par la fonction de transport T. Dans ce travail, nous proposons de choisir $d(\cdot, \cdot)$ comme la distance de Wasserstein $W_p(\cdot, \cdot)$. Ce choix sera motivé par le fait que cette distance peut être approchée aisément sans connaissance explicite des lois de probabilités μ et ν lorsque celles-ci sont décrites empiriquement par des échantillons. La relaxation du problème de Monge (1) pourra alors s'écrire

$$\inf_T \left\{ W_p(T_\sharp\mu, \nu) + \lambda \int_{\mathcal{X}} c(x, T(x))\mathrm{d}\mu(x) \right\} \tag{2}$$

où la fonction de coût définie dans (1) est interprétée ici comme un terme de régularisation.

Remarque *La formulation relaxée* (2) *fait apparaître la distance de Wasserstein entre la mesure cible ν et la mesure image $T_\sharp\mu$. Ce terme ne doit pas être confondu avec la distance de Wasserstein $W_p(\mu, \nu)$ qui est l'infimum atteint par la solution du problème de transport au sens de Kantorovitch.*

2.3 Discrétisation du problème

Dans un contexte d'apprentissage automatique, nous disposons d'échantillons qui permettent d'approcher les mesures continues sous-jacentes par des mesures ponctuelles empiriques. Dans cet article, nous nous intéressons donc aux mesures discrètes et à la formulation empirique du problème de transport optimal. Ainsi, dans un cadre discret, on considérera alors μ et ν deux mesures discrètes décrites par les échantillons respectifs $\mathbf{x} = \{x_n\}_{n=1}^N$ et $\mathbf{y} = \{y_n\}_{n=1}^N$ telles que $\mu = \frac{1}{N}\sum_{n=1}^N \delta_{x_n}$ et $\nu = \frac{1}{N}\sum_{n=1}^N \delta_{y_n}$. Dans ce qui suit, nous proposons une version empirique du critère (2) dans le cas de mesures discrètes.

La formulation (2) requiert d'évaluer une distance de Wasserstein dont le calcul n'est pas trivial sous sa forme originale notamment en grande dimension. Une alternative consiste à considérer sa réécriture sous la forme de la *sliced-Wasserstein* (SW). L'idée sous-jacente à la distance SW consiste à représenter une distribution définie en grande dimension à l'aide un ensemble de distributions unidimensionnelles pour lesquels le calcul de la distance de Wasserstein est explicite [5]. Notons p_X et p_Y les distributions de probabilité des variables aléatoires X et Y. Pour tout vecteur de la sphère unité $u \in \mathbb{S}^{d-1}$ nous définissons l'opérateur de projection $S_u : \mathbb{R}^d \to \mathbb{R}$ par $S_u(x) \triangleq \langle u, x \rangle$. La distance SW d'ordre $p \in [1, \infty)$ entre p_X et p_Y peut s'écrire [5]

$$SW_p\left(p_X, p_Y\right) = \left(\int_{\mathbb{S}^{d-1}} W_p\left(S_{u\sharp}p_X, S_{u\sharp}p_Y\right)^p du \right)^{\frac{1}{p}} \tag{3}$$

où la distance $W_p(\cdot, \cdot)$ définissant l'intégrande est unidimensionnelle, conduisant à un calcul explicite par inversion des fonctions de répartition. Dans le cas où les distributions p_X et p_Y sont représentées par les échantillons respectifs $\mathbf{x}$ et $\mathbf{y}$, une approximation de Monte Carlo de la distance SW est

$$\widehat{SW}_p(\mathbf{x}, \mathbf{y}) = \frac{1}{J}\sum_{j=1}^J W_p\left(\frac{1}{N}\sum_{n=1}^N \delta_{S_{u_j}(x_n)}, \frac{1}{N}\sum_{n=1}^N \delta_{S_{u_j}(y_n)} \right)$$

où $u_1, \ldots, u_J$ sont tirés uniformément sur la sphère $\mathbb{S}^{d-1}$. La forme empirique de la relaxation du problème de Monge (2)

s'écrit alors

$$\min_{T} \left\{ \widehat{SW}_p(T(\mathbf{x}), \mathbf{y}) + \lambda \sum_{n=1}^{N} c\left(x_n, T\left(x_n\right)\right) \right\}. \quad (4)$$

3 Flux de normalisation pour le transport optimal

Cette section propose de résoudre le problème (4) en restreignant la classe des opérateurs T à une famille de réseaux profonds inversibles dénommés flux de normalisation. La structure et les principales propriétés de ces réseaux sont détaillés au paragraphe 3.1. La stratégie d'apprentissage proposée permettant d'entraîner ces réseaux pour résoudre le problème (4) est alors détaillée au paragraphe 3.2.

3.1 Flux de normalisation

Les flux de normalisation sont une classe flexible de réseaux génératifs profonds qui cherchent à apprendre un changement de variable entre deux distributions p_X et p_Y grâce à une transformation inversible $T_{\boldsymbol{\theta}} : X \mapsto Y = T_{\boldsymbol{\theta}}(X)$. Habituellement, la distribution p_X n'est connue qu'à travers des échantillons $\mathbf{x} = \{x_n\}_{n=1}^{N}$ et, par simplicité, la distribution p_Y est choisie comme une loi normale centrée réduite. Les paramètres $\boldsymbol{\theta}$ définissant l'opérateur $T_{\boldsymbol{\theta}}$ sont alors ajustés par maximisation de la vraisemblance associée aux observations $\mathbf{x}$ en utilisant

$$p_X(x) = p_Y\left(T_{\boldsymbol{\theta}}(x)\right) \left|\det J_{T_{\boldsymbol{\theta}}^{-1}}\right| \quad (5)$$

où $J_{T_{\boldsymbol{\theta}}^{-1}} = \frac{\partial T_{\boldsymbol{\theta}}^{-1}}{\partial x}$. Dans ce travail, nous considérons des réseaux $T_{\boldsymbol{\theta}} = T_{\boldsymbol{\theta}_M}^{(M)} \circ T_{\boldsymbol{\theta}_{M-1}}^{(M-1)} \circ \ldots \circ T_{\boldsymbol{\theta}_1}^{(1)}$ composés de M couches de couplage affine qui assurent une transformation inversible et une expression explicite du jacobien, avec $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$. Dans la suite, pour alléger les notations, nous écrirons $T_{\boldsymbol{\theta}_m}^{(m)} = T_m$. L'entrée et la sortie de la m-ième couche sont définies par la relation

$$\left(y_{\text{id}}\,, y_{\text{ch}}\right) = T_m\left(x_{\text{id}}\,, x_{\text{ch}}\right)$$

avec

$$\begin{cases} y_{\text{id}} = x_{\text{id}} \\ y_{\text{ch}} = \left(x_{\text{ch}} + \mathsf{D}_m\left(x_{\text{id}}\right)\right) \odot \exp\left(\mathsf{E}_m\left(x_{\text{id}}\right)\right) \end{cases} \quad (6)$$

où x_{id} et x_{ch} (resp. y_{id} et y_{ch}) sont des sous-ensembles disjoints de composantes du vecteur d'entrée x (resp. du vecteur de sortie y). La séparation de l'entrée x en x_{id} et x_{ch} s'effectue par masquage où seule la partie $x_{\text{ch}} = \text{mask}(x)$ est transformée en fonction de la partie inchangée x_{id}. Les fonctions d'échelle $\mathsf{E}_m(\cdot)$ et de décalage $\mathsf{D}_m(\cdot)$ sont alors décrites par des réseaux de neurones dont les paramètres $\boldsymbol{\theta}_m$ sont à ajuster. Parmi les exemples de flux avec des couches de couplage, nous pouvons citer RealNVP [3] et Glow [4]. Cette spécificité architecturale permet d'accéder à une discrétisation de la fonction de transport $T_{\boldsymbol{\theta}}$ en une composition de fonctions de sous-transport T_m.

3.2 Fonction de coût pour l'entraînement

Comme indiqué précédemment, l'objectif de ce travail vise à apprendre un opérateur bijectif permettant d'associer deux distributions quelconques p_X et p_Y pour lesquelles nous avons accès uniquement à des échantillons $\mathbf{x}$ et $\mathbf{y}$. Nous restreignons la recherche de cet opérateur à la classe des réseaux profonds inversibles $T_{\boldsymbol{\theta}}$ décrits au paragraphe 3.1. La stratégie conventionnelle d'apprentissage par maximisation de la vraisemblance (5) ne peut être mise en oeuvre ici puisque la distribution de base p_Y n'est plus explicitement donnée mais n'est accessible qu'à travers la connaissance de l'échantillon $\mathbf{y}$. Pour ajuster les poids $\boldsymbol{\theta}$ du réseau, une alternative consiste à interpréter la tâche sous-jacente d'apprentissage comme la recherche d'une fonction de transport. Une première approche consisterait à ajuster ces poids en résolvant directement le problème (4). Cependant, pour tirer parti de l'architecture en couches de l'opérateur $T_{\boldsymbol{\theta}}$ recherché, il apparaît légitime d'essayer de répartir les efforts de transport assurés par chaque couche. La régularisation apparaissant dans (4) sera donc instanciée pour chaque transformation élémentaire T_m réalisée par chaque couche du réseau.

Par ailleurs, un défi majeur lors de l'ajustement des modèles d'apprentissage profond est l'obligation de se limiter à des informations partielles pour déduire la structure globale du paysage d'optimisation. De plus, dans notre cas, la fonction de coût à optimiser n'est pas constante puisque l'approximation $\widehat{SW}$ de la distance SW intervenant dans (4) dépend du tirage de vecteurs unitaires $\{u_j\}_{j=1}^{J}$. Afin d'atténuer ces difficultés d'optimisation, nous proposons d'introduire des pénalisations complémentaires $\left|J_{T_m}(\cdot)\right|^2$ sur les jacobiens des transformations T_m. Ces pénalisations de type Sobolev promeuvent des opérateurs réguliers T_m, assurant également un opérateur global $T_{\boldsymbol{\theta}}$ lui-même régulier. Dans le cadre du transport optimal, cette régularisation a déjà été étudiée dans [6] où il est montré que ce type de pénalisation de la formulation de Monge permet de tendre vers un opérateur optimal au sens du transport. Finalement, l'entraînement du réseau est réalisé en minisant la fonction de coût

$$\sum_{n=1}^{N} \sum_{m=1}^{M} \Bigg[\lambda c(T_{m-1}(x_n), T_m(x_n)) + \gamma \left|J_{T_m}(x_n)\right|^2 \Bigg] + \widehat{SW}_p(\mathbf{x}, \mathbf{y}) \quad (7)$$

avec $T_0(x_n) = x_n$. La fonction de transport obtenue est une composition de fonctions de sous-transports. La mesure image par l'opérateur $T_{[m]} \triangleq T_m \circ \ldots \circ T_1$ en sortie de la $m$ième couche ($m = 1, \ldots, M$ avec $T_{[M]} \triangleq T_{\boldsymbol{\theta}}$) peut être interprétée comme un barycentre de Wasserstein entre μ et ν [7]. Avec $\alpha_m = \frac{m}{M}$, la mesure $T_{[m]\sharp}\mu$ peut être interprétée comme la solution du problème

$$\inf_{\beta} \left\{ \alpha_m W_p(\mu, \beta) + (1 - \alpha_m) W_p(\nu, \beta) \right\}. \quad (8)$$

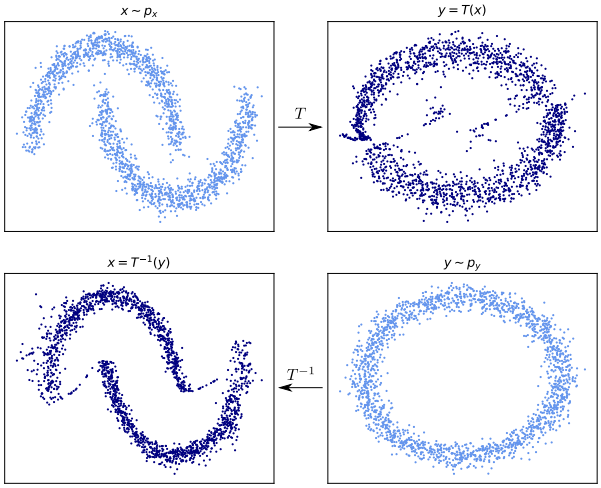

FIGURE 1 – Résultats obtenus lorsque la distribution cible p_X est une double-lune (en haut à gauche) et la distribution de base p_Y est un cercle (en bas à droite).

4 Expériences et résultats

Cette partie illustre la méthode proposée à l'aide d'expériences numériques réalisées sur des jeux de données synthétiques. Pour les expériences ci-dessous, nous utiliserons la distance euclidienne comme fonction du coût du transport, i.e., $c(a, b) = \|a - b\|_2^2$. L'opérateur recherché $T_{\boldsymbol{\theta}}$ est choisi comme un réseau d'architecture de type RealNVP [3] de 4 flux composés chacun de deux réseaux de neurones à quatre couches, correspondant aux fonctions $\mathsf{E}_m(\cdot)$ et $\mathsf{D}_m(\cdot)$ dans (6). L'entrainement est réalisé par optimisation sur des mini-lots de 1000 échantillons avec un optimiseur Adam pour un nombre total de 20000 échantillons par mesure de probabilité.

La Fig. 1 représente les résultats obtenus après apprentissage d'un opérateur T qui transporte une distribution p_X en forme de double lune (en haut à gauche) vers une distribution en forme de cercle (en bas à droite). Sont également représentées les mesures images empiriques $T_\sharp p_X$ (en haut à droite) et $T_\sharp^{-1} p_Y$ (en bas à gauche) qui sont obtenues par application de l'opérateur $T(\cdot)$ estimé ou de son inverse $T^{-1}(\cdot)$.

La Fig. 2 vise à illustrer l'intérêt de la régularisation de type Sobolev (i.e., norme du Jacobien) dans la fonction de coût (7). Dans cette expérience, l'objectif est d'apprendre la fonction de transport entre une distribution en cercle p_X (bleu clair) vers une autre distribution p_Y en cercle qui est translatée (bleu foncé). Sont représentées sur cette figure les sorties de chacune des M couches du réseau, c'est-à-dire les mesures images $T_{[m]\sharp} p_X$ pour $m = 1, \ldots, M$. En l'absence de régularisation (à gauche), les résultats successifs des sous-transports souffrent clairement de multiples déformations superflues (translations et dilatations). Lorsque la fonction de coût est munie d'une pénalisation de type Sobolev (à droite), l'opérateur T appris de décompose comme une succession de sous-transports beaucoup plus réguliers. Ceci est confirmé par le coût total du transport qui est de 360 avec régularisation contre 520 sinon. De plus ce

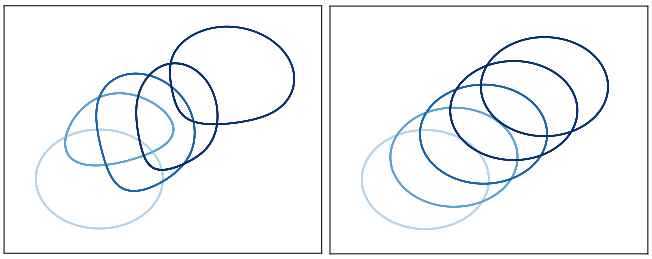

FIGURE 2 – Trajectoires des transports appris sans régularisation de type Sobolev (à g.) ou avec régularisation (à dr.).

coût est réparti de façon homogène entre les couches successives, avec une variation d'au plus $\pm 1\%$ d'une couche à l'autre, contre $\pm 20\%$ sinon.

5 Conclusion

La contribution principale de cet article est une utilisation des flux de normalisation pour apprendre une approximation d'une fonction de transport optimal entre deux distributions empiriques. Nous proposons une formulation relaxée et pénalisée du problème de Monge. Cette formulation est utilisée comme fonction de coût pour entraîner un réseau bijectif dont l'architecture permet l'accès à des transports intermédiaires, pouvant être associés à des barycentres de Wasserstein. La méthode proposée est illustrée par des expériences numériques sur des exemples-jouets. Les travaux futurs viseront à étendre ces résultats à des expériences en grande dimension.

Remerciements

Les auteurs remercient Elsa Cazelles (IRIT, CNRS UMR 5505) et Benoît Merlet (Laboratoire Paul Painlevé, CNRS UMR 8524) pour les discussions qui ont ponctué ce travail.

Références

[1] G. Monge, "Mémoire sur la théorie des déblais et des remblais," *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[2] G. Peyré and M. Cuturi, "Computational optimal transport : with applications to data science," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[3] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," in *Proc. IEEE Int. Conf. Learn. Represent. (ICLR)*, 2017.

[4] D. P. Kingma and P. Dhariwal, "Glow : Generative flow with invertible 1x1 convolutions," in *Adv. in Neural Information Process. Systems (NeurIPS)*, 2018.

[5] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister, "Sliced and radon Wasserstein barycenters of measures," *J. Math. Imag. Vision*, vol. 51, no. 1, pp. 22–45, 2015.

[6] J. Louet, "Optimal transport problems with gradient penalization," Ph.D. dissertation, Université Paris Sud-Paris XI, 2014.

[7] M. Agueh and G. Carlier, "Barycenters in the Wasserstein space," *SIAM J. Mathematical Analysis*, vol. 43, no. 2, pp. 904–924, 2011.