INTRODUCTION AU DOMAINE DE RECHERCHE : ESPACE DE WASSERSTEIN ET ENTROPIE

HUGO MALAMUT

AVANT-PROPOS

L'objectif de ce document est de donner une introduction au domaine du transport optimal et ses interactions avec la fonctionnelle d'entropie. La recherche dans le domaine porte plutôt sur des questions d'analyse, mais ici le point de vue adopté est le plus géométrique possible. L'accent est mis sur les concepts et les idées davantage que sur la rigueur mathématiques.[1] Enfin, bien que la langue du domaine soit l'anglais, ce document est rédigé en français. D'où de possibles traductions mot-à-mot de l'anglais dont je m'excuse par avance. Bonne lecture !

NOTATIONS

- M : *espace de base*, variété de dimension finie, souvent $\mathbb{R}^d$ ou le tore $\mathbb{R}^d/\mathbb{Z}^d$,
- x, y désigneront des points de M : $x \in M$, $y \in M$
- T_xM plan tangeant à M en x
- TM fibré tangeant de M
- $(x, \vec{v})$ désignera un vecteur tangeant à M en x
- d : dimension de M *ou* distance sur M (euclidienne pour $\mathbb{R}^d$, euclidienne périodique pour le tore)
- T : application de M dans M
- $\mathcal{C}(M)$: fonctions continues de M dans $\mathbb{R}$
- $\mathcal{P}(M)$: ensembles des mesures de probabilité sur M
- μ, ν désigneront des mesures sur M : $\mu \in \mathcal{P}(M)$, $\nu \in \mathcal{P}(M)$
- X,Y : désignent des variables aléatoire à valeur dans M
- $X \sim \mu$: X a comme loi μ
- $\mathcal{P}^2(M)$ mesures admetant un moment d'ordre 2 quand cela a un sens ($M = \mathbb{R}^d$). Pour le tore, $\mathcal{P}^2(M) = \mathcal{P}(M)$
- $\mathcal{L}$, dx : mesure de Lebesgue sur M
- $\mathcal{P}_a(M)$ mesures absolument continues par rapport à $\mathcal{L}$
- T_μ : poussé en avant de μ par T[2]
- δ_x : mesure de Dirac au point x
- f : opérateur monotone, $\forall x, y \in \mathbb{R}^d \quad \langle f(x) - f(y), x - y \rangle \geq 0$
- c : fonctionnelle de coût, généralement $c(x, y) = \|x - y\|^2$
- $(\mu_t)_{t\in\mathbb{R}}$: un chemin dans $\mathcal{P}(M)$
- $(\rho_t)_{t\in\mathbb{R}}$: lorsque μ_t est absolument continue, ρ_t est sa densité, $d\mu_t(x) = \rho_t(x)dx$
- ∇, div : opérateurs gradient et divergence
- H : fonction sur $\mathcal{P}(M)$: $H : \mathcal{P}(M) \to \mathbb{R}$

Date: May 2022.

[1] Il est même possible que des passages soient faux

[2] $\forall f : M \mapsto \mathbb{R}$ *mesurable, bornée*, $\int_M f \mathrm{d}(T_\# \mu) = \int_M f \circ T \mathrm{d}\mu$

CONTENTS

INTRODUCTION

A propos de transport optimal et d'entropie, trois sujets naturels viennent à l'esprit :

- la régularisation entropique du transport optimal et l'algorithme de Sinkhorn.
- la théorie synthétique de la courbure de Ricci
- l'équation de diffusion vue comme flot gradient dans l'espace de Wasserstein

Par manque de place pour le premier et de connaissances pour le second, j'ai décidé de ne présenter que le troisième. Ce document est organisé comme suit : Dans une première partie, le domaine du transport optimal est présenté rapidement. Dans une seconde partie, la notion de gradient Wasserstein est introduite, afin d'expliquer en quoi "l'équation de diffusion est le flot gradient de l'entropie dans l'espace de Wasserstein". Enfin, la dernière partie présente une liste de questions personnelles sur le comportement asymptotique de l'entropie au voisinage de mesures concentrées.

1. TRANSPORT OPTIMAL, ESPACE DE WASSERSTEIN

1.1. **Problème de Monge, formulation probabiliste. A l'origine du transport optimal** se trouve une question posée en premier par Gaspard Monge au 18ème siècle pour $M = \mathbb{R}^d$: considérant deux distributions de masses :

- une distribution *de départ*, appellée μ dans la suite et représentée en bleu
- une distribution *d'arrivée*, appellée ν dans la suite et représentée en rouge

Et en supposant que la masse totale est la même (normalisée à 1 dans la suite), comment transporter la masse de façon *la plus efficace possible* ? En termes modernes, Gaspard Monge cherche une application T telle que $T_{\#}\mu = \nu$ (voir figure 1) qui minimise $\int_M d(x, T(x))d\mu(x)$. Par la suite historiquement, la distance d va être remplacé par toute sorte de coût, et en particulier le coût quadratique va s'imposer, d'où la définition suivante :

Définition 1 (Problème du transport optimal)**.** Soit $\mu, \nu \in \mathcal{P}(M)$. Le problème de Monge associé à μ et ν est : trouver une application $T^* : M \mapsto M$ telle que

$$T^* := \underset{\nu = T_{\#}\mu}{\operatorname{argmin}} \int_M d(x, T(x))^2 \, \mathrm{d}\mu(x) \tag{M}$$

Quand elle existe, T^* est un *transport optimal* entre μ et ν. On parle aussi d'application de Brenier

Remarque 1 (Théorème de Brénier et opérateurs monotones[a]). *Dans le cas $M = \mathbb{R}^d$, il y a un lien très intéressant avec les opérateurs monotones : on peut développer $d(x, T(x))^2 = \|x - T(x)\|^2$, si bien que $\int_M d(x, T(x))^2 d\mu(x) = M_2(\mu) + M_2(\nu) - 2\int_M \langle x, T(x)\rangle d\mu(x)$, avec M_2 le moment d'ordre 2. Donc si T^* est un transport optimal, T maximise*

$$\int_M \langle x, T(x)\rangle d\mu(x)$$

parmi les transports de μ à ν. D'un autre côté, si f est un opérateur monotone, on peut réecrire la propriété $\langle f(x) - f(y), x - y\rangle \geq 0$ comme $\frac{1}{2}\langle f(x), x\rangle + \frac{1}{2}\langle f(y), y\rangle \geq \frac{1}{2}\langle f(x), y\rangle + \frac{1}{2}\langle f(y), x\rangle$, autrement dit, f est un opérateur monotone si et seulement si f est un transport optimal de μ à $f_{\#}\mu$ pour toute mesure μ de la forme $\mu = \frac{1}{2}\delta_x + \frac{1}{2}\delta_y$. De manière analogue, Brenier a découvert que f est un gradient de fonction convexe si et seulement si f est un transport optimal de μ à $f_{\#}\mu$ pour toute mesure μ.

[a] $f : \mathbb{R}^d \to \mathbb{R}^d$ est monotone si $\langle f(x) - f(y), x - y\rangle \geq 0 \quad \forall x, y$

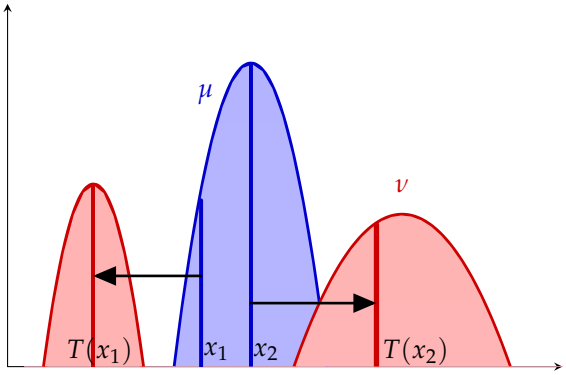

FIGURE 1. l'application T transporte μ vers ν : $T_{\#}\mu = \nu$

Formulation probabiliste Le problème de Monge est assez mal posé, notamment puisqu'il n'existe pas toujours d'application T telle que $T_{\#}\mu = \nu$. Dans les années 30, Kantorovich va proposer une relaxation du problème qui permet qu'il soit toujours bien posé et même qu'il existe toujours une solution (pas unique tout de même).

Définition 2 (Problème de Kantorovich). Soit $\mu, \nu \in \mathcal{P}(M)$. Le problème de Kantorovich associé à μ et ν est :

$$\inf_{X\sim\mu Y\sim\nu} \mathbb{E}[d(X, Y)^2]$$

Lorsqu'un minimiseur (X^*, Y^*) existe, on parle de *couplage optimal* et on note π^* la loi du couple (X^*, Y^*). π^* est appelé *plan de transport optimal* entre μ et ν

Ce problème a toujours des solutions, et la valeur du problème est très intéressante, car elle permet de définir une distance sur l'espace $\mathcal{P}^2(M)$

Définition 3 (Théorème). Soit $\mu, \nu \in \mathcal{P}^2(M)$, La *distance de Wasserstein* de μ à ν, $W_2(\mu, \nu)$ est définie par

$$W_2^2(\mu, \nu) := \min_{X\sim\mu Y\sim\nu} \mathbb{E}[d(X, Y)^2]$$

La distance de Wasserstein est une distance sur $\mathcal{P}^2(M)$ qui métrise la topologie faible-*

1.2. **Dualité de Kantorovich.** En introduisant la relaxation ci-dessus, Kantorovich a remarqué que le nouveau problème était désormais linéaire en π le plan de transport : il s'agit de minimiser $\int c d\pi$ avec c le coût quadratique $c(x,y) := \|x-y\|^2$ et sous la contrainte que μ et ν soient les marginales de π. Ce problème admet un problème dual naturel, en écrivant la contrainte de marginale $\int_M f(x)d\pi(x,y) - \int f\mu = 0 \quad \forall f \in \mathcal{C}(M)$, et en utilisant une méthode de dualité convexe, on obtient le problème dual suivant :

Proposition 1 (Problème Dual)**.** *Dans le cas $M = \mathbb{R}^d$ et $M = \mathbb{R}^d/\mathbb{Z}^d$, on a*

$$W_2^2(\mu,\nu) = \max_{\substack{f,g\in\mathcal{C}(M)\\ f+g\leq c}} \int f d\mu + \int g d\nu$$

où la contrainte $f+g \leq c$ signifie $f(x)+g(y) \leq \|x-y\|^2 \quad \forall x,y$. Un couple (f,g) optimal constitue une paire de potentiels de Kantorovich

Remarque 2 (Problèmes convexes et dualité)**.** *Soit X un espace, f une fonction qu'on tente de minimiser sur $x \in X$ sous la contrainte $g(x) = 0\, \forall g \in V$ avec V un $\mathbb{R}$-espace vectoriel. Comme V est un espace vectoriel, $\max_{g\in V} g(x)$ vaut 0 si tous les $(g(x))_{g\in V}$ sont nuls, et $+\infty$ autrement. On a donc*

$$\min_{\forall g\in V g(x)=0} f(x) = \min_x (f(x) + \max_{g\in V} g(x)) = \min_{x\in X} \max_{g\in V} (f(x)+g(x))$$

Le problème $\max_{g\in V} \min_{x\in X} f(x)+g(x)$ est appelé problème dual *: le* min max *est toujours plus grand que le* max min *donc ce problème dual a une valeur plus petite que celle du* problème primal*.*

Théorème 2 (Brenier)**.** *Si $\mu \in \mathcal{P}_a(M)$, alors il existe une application de transport optimal T*[3] *En outre on a aussi existence d'un plan optimal π et de potentiels de Kantorovich (f,g). Les liens entre ces objets sont les suivants :*

- $\nabla f = 2(T - id)$,
- *π est concentré sur le graphe de T.*

2. La structure riemannienne de l'espace de Wasserstein

2.1. **vecteurs aléatoires et structure riemanienne.** La formulation de Kantorovich a l'avantage d'être symétrique en μ et ν. Mais cette symétrie est aussi trompeuse, et il peut être intéressant de considérer le problème non pas sur les couples (X,Y) de variables aléatoires à valeur dans $M \times M$, mais sur les couples $(X,\vec{v})$ à valeurs dans TM, avec le changement de variable $\vec{v} = \vec{XY}$ (pour $\mathbb{R}^d$)

Définition 4 (Proposition)**.** Soit μ, ν deux mesures de probabilité. Le problème de transport optimal correspond à

$$W_2^2(\mu,\nu) = \min_{\substack{X\sim\mu\\ X+\vec{v}\sim\nu}} \mathbb{E}[\|\vec{v}\|^2]$$

Un couple $(X,\vec{v})$ est appelé *vecteur optimal*, sa loi sera notée γ^* et aussi appelée vecteur optimal de manière abusive.

Dans ce cadre, on sent bien que l'espace de Wasserstein a une structure riemanienne. Les lois γ de variables $(X,\vec{v})$ sur TM qui ont comme première marginale μ et avec $\mathbb{E}[\|\vec{v}\|^2] < +\infty$ sont les vecteurs tangents à μ. Bien sûr, la norme de γ est donnée par $\mathbb{E}[\|\vec{v}\|^2]$. On peut aller un peu plus loin en définissant un produit scalaire : si γ_1 et γ_2 sont deux vecteurs tangents en μ,

$$\langle \gamma_1, \gamma_2 \rangle_\mu := \sup_{\substack{(X,\vec{v}_1)\sim\gamma_1\\ (X,\vec{v}_2)\sim\gamma_2}} \mathbb{E}(\langle v_1, v_2 \rangle)$$

[3] et si $M = \mathbb{R}^d$, alors T est le gradient d'une fonction convexe

Avant de continuer, il est intéressant de distinguer parmi les vecteurs aléatoires ceux qui sont en réalité déterministes.

Définition 5. Soit $(X, \vec{v}) \sim \gamma$ un vecteur aléatoire entre μ et ν. ($X \sim \mu$ et $X + \vec{v} \sim \nu$) Le vecteur est dit *déterministe* si γ est la loi d'un couple $(X, \vec{v}(X))$. Si γ est déterministe, il est déterminé par l'application $\vec{v}$.

Si γ_1 et γ_2 sont tous deux déterministes, correspondant à $\vec{v}_1$ et $\vec{v}_2$, alors

$$\langle \gamma_1, \gamma_2 \rangle_\mu = \int \langle \vec{v}_1(x), \vec{v}_2(x) \rangle d\mu(x)$$

Pour les vecteurs optimaux, la situation déterministe correspond à l'existence d'une application optimale, comme illustré par la figure 2. Lorsque $\mu \in \mathcal{P}_a(M)$, c'est toujours le cas. La figure 3 illustre que $\mu \notin \mathcal{P}_a(M)$ rend souvent nécessaire que le vecteur $\vec{v}$ soit réellement aléatoire.

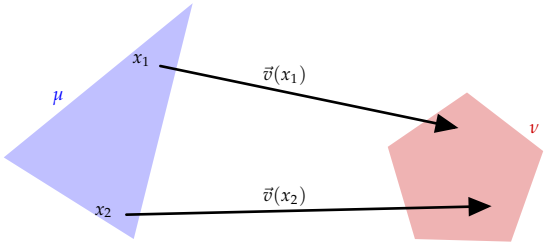

FIGURE 2. Le vecteur $\vec{v}$ est déterministe au sens où il est fixé par x : cette situation correspond à l'existence d'une application optimale

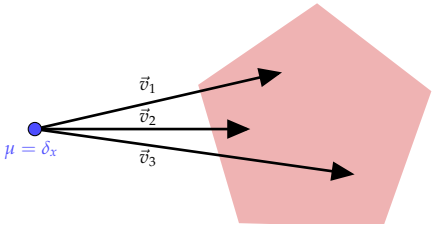

FIGURE 3. Le vecteur aléatoire $\vec{v}$ prend plusieurs valeurs (dont v_1,v_2 et v_3) alors que x est fixé

2.2. **Chemins et vecteurs dérivés.** Le but de cette partie est de comprendre en quoi l'équation de diffusion $\partial_t \rho = \Delta \rho$ est un flot gradien Wasserstein. Cette équation est écrite dans un formalisme eulerien. Or l'espace de Wasserstein correspond davantage à une vision lagrangienne (particules qui bougent dans M), tandis que le formalisme eulerien correspond davantage à l'espace $L^2(\mathcal{L})$ (densités qui montent ou descendent). Il est donc nécessaire de faire le lien entre ces deux espaces.

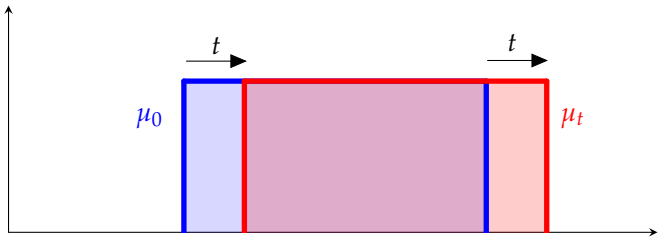

FIGURE 4. La variation verticale est trop rapide : le chemin μ_t est dérivable pour la structure W_2 mais pas pour la structure $\mathcal{L}^2$

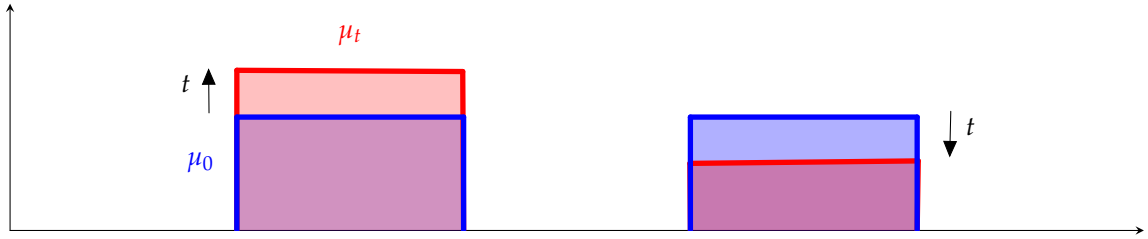

FIGURE 5. La variation horizontale est trop rapide : le chemin μ_t est dérivable pour la structure L^2, mais pas pour la structure W_2

Vision lagrangienne et eulerienne Lorsqu'un fluide se déplace, on peut étiqueter ses particules avec un espace Ω et noter $X_t(\omega)$ la position de la particule ω au temps t. L'équation du mouvement donne une équation différentielle sur X_t, faisant éventuellement intervenir la distribution de X_t (zones plus concentrées, moins concentrées...). Cette équation différentielle d'évolution est respectée pour tout ω. C'est la *vision lagrangienne*. Pour que cela fonctionne, il faut doter Ω d'une mesure $\mathcal{L}$ telle que la loi de X_0, $(X_0)_\#\mathcal{L}$ corresponde à la concentration initial du fluide. On peut également se concentrer uniquement sur la loi de X_t, μ_t, et, si cette loi est assez régulière, disons $\mu_t \in \mathcal{P}_a(M)$, obtenir une équation différentielle sur ρ_t, la densité de μ_t. C'est le *point de vue eulerien*. Dans un cas on regarde des variables aléatoires, dans l'autre des mesures de probabilité.

Comme on s'intéresse à $L^2(\mathcal{L})$, avec $\mathcal{L}$ la mesure de Lebesgue, on va considérer uniquement des mesures absolument continues. On notera ρ_t la densité de μ_t. Dans ce cas, grâce au théorème de Brenier, on peut se limiter à des vecteurs déterministes dans l'espace de Wasserstein.

A propos des différences entre structure L^2 et W_2, il faut tout d'abord avoir conscience que certains chemins $(\mu_t)_{t\in\mathbb{R}}$ dans $\mathcal{P}_a(M)$ sont dérivables pour W_2 mais pas pour L^2 et vice-versa. Les figures 4, 5 et 6 en sont des illustrations.

Si (μ_t) est dérivable pour la structure $\mathcal{L}^2(\mathcal{L})$, sa dérivée s'aparrente à une mesure signée d'intégrale nulle notée $\partial_t\rho_t$. Si (μ_t) est W_2-dérivable avec un vecteur déterministe, on note $\vec{v}_t$ le vecteur dérivé.

Le lien entre les deux est donné par l'équation de continuité

$$\partial_t\rho_t = -\operatorname{div}(\rho_t\vec{v}_t)$$

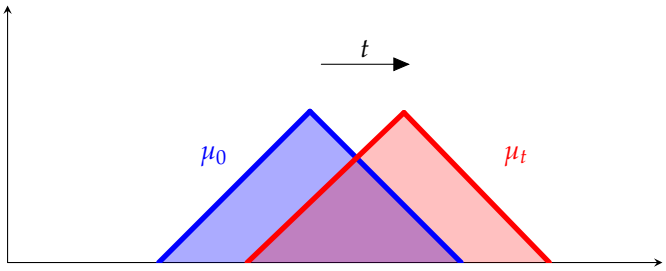

FIGURE 6. Les variations horizontales et verticales sont lentes : μ_t est dérivable pour les deux structures $\mathcal{L}^2$ et W_2

L'équation de continuité Il n'est pas très surprenant que l'équation de continuité fasse le lien entre point de vue eulerien et lagrangien. On peut le voir "à la physicienne" avec le théorème de Stokes : pour un petite partie $A \subset M$, si le fluide de densité ρ_t est animé par la vitesse v_t, $\int_A \operatorname{div}(\rho_t \vec{v}_t) dx$ correspond au flux sortant de A, qui doit être l'opposé de la variation de masse dans A, $\frac{d}{dt}\int_A \rho_t dx = \int_A \partial_t \rho_t dx$. Pour A tendant vers un point $x \in M$ quelconque, $\partial_t \rho_t(x) = -\operatorname{div}(\rho_t(x)\vec{v}_t(x))$
On peut aussi utiliser un argument d'analyse fonctionnelle. Soit f une fonction test. $\vec{v}_t$ est la dérivée particulaire (lagragienne) de μ_t , donc $\mu_{t+dt} \simeq (x + dt\vec{v}_t(x))_{\#}\mu_t$, où x est la fonction identité. donc au premier ordre, $\int f d\mu_{t+dt} \simeq \int f(x + dt\vec{v}_t(x)) d\mu_t(x) \simeq \int f(x) d\mu_t(x) + dt \int \nabla f(x) \cdot \vec{v}_t(x) \rho_t(x) dx$.
Mais $\int \nabla f \cdot \vec{v}_t \rho_t dx = -\int f \operatorname{div}(\rho_t \vec{v}_t) dx$[a] On a donc

$$\int f d\mu_{t+dt} \simeq \int f d\mu_t - dt \int f \operatorname{div}(\rho_t \vec{v}_t) dx$$

Mais comme $\partial_t \rho_t$ est la dérivée $L^2(\mathcal{L})$ de μ_t,

$$\int f d\mu_{t+dt} \simeq \int f d\mu_t + dt \int \partial_{\rho_t} f dx$$

On peut donc identifier $\int \partial_{\rho_t} f dx = -\int f \operatorname{div}(\rho_t \vec{v}_t) dx$ pour toute fonction test f : $\partial \rho_t = -\operatorname{div}(\rho_t \vec{v}_t)$.

[a] M est sans bord, donc $-$ div est l'*opérateur adjoint* du gradient, car $0 = \int_M \operatorname{div}(f\vec{v}) = \int f \operatorname{div}(v) + \int \nabla f \cdot \vec{v}$

Pour récapituler, pour les vecteurs déterministes, selon que les mesures μ_t sont absolument continues ou empiriques[4], on a deux façon d'exprimer l'équation $\vec{v}_t = \left(\frac{d\mu_t}{dt}\right)_W$:

Mesures empiriques	Mesures absoluments continues
$\mu_t = \sum_{i=1}^{N} a_i \delta_{x_i(t)}$	$d\mu_t(x) = \rho_t(x) dx$
$\frac{d}{dt} x_i = \vec{v}_t(x_i)$	$\partial_t \rho_t = -\operatorname{div}(\rho_t \vec{v}_t)$

2.3. **Gradients L^2 et Wasserstein.** Maintenant que nous avons étudié en détail les chemins sur $(\mathcal{P}(M), W_2)$, on peut s'intéresser à la notion duale : les fonctions sur $(\mathcal{P}(M), W_2)$. Soit H une fonction sur $\mathcal{P}(M)$

$$\begin{array}{ccccc} \mathbb{R} & \to & \mathcal{P}_2(M) & \to & \mathbb{R} \\ t & \mapsto & \mu_t & & \\ & & \mu & \mapsto & H(\mu) \end{array}$$

[4] Une mesure empirique est une mesure constituée de Diracs

On va s'intéresser aux différents gradients de H selon la structure riemanienne.

Structure riemanienne et gradient Soit E une véritable variété riemanienne (de dimension finie). Chaque plan tangent T_xE est doté d'un produit scalaire $\langle , \rangle_x$.
Le gradient $\nabla_x H$ est défini par représentation de Riesz de la différentielle $d_x H$ dans $(T_xE, \langle , \rangle_x)$, ie par la formule :

$$\forall (x, \vec{v}) \in TE \quad d_x H \cdot \vec{v} = \langle \nabla_x H, \vec{v} \rangle_x$$

Si (x_t) est un chemin dans E de dérivée (v_t), on a alors

$$\frac{d}{dt} H(x_t) = \langle \nabla_{x_t} H, \vec{v}_t \rangle_{x_t}$$

Et cette formule caractérise également le gradient.

On notera $\nabla^{L^2} H$ le gradient de H dans $L^2(\mathcal{L})$, et $\nabla^W H$ le gradient dans W_2.

Exemple 3 (de gradients L^2)**.**

(1) *Si* $H(\mu) = \int f d\mu$, *et* $d\mu_t = \rho_t dx$, $\frac{d}{dt}\left(\int f \rho_t dx\right) = \int f \partial_t \rho_t dx$ *Et comme comme* $\partial_t \rho_t$ *est la dérivée de* μ_t *dans* $L^2(\mathcal{L})$, $\nabla^{L^2} H = f$
(2) *Si H est définie sur* $\mathcal{P}_a(M)$ *par* $H(\rho) = \int g(\rho) dx$, *alors de même* $\frac{d}{dt} H(\rho_t) = \int \partial_t (g(\rho_t)) dx = \int g'(\rho_t) \partial_t \rho_t dx$ *et donc* $\nabla^{L^2}_\rho H = g'(\rho)$
(3) *En applicant ce dernier exemple pour* $g(x) = x \ln(x)$, *on a le gradient* L^2 *de la fonctionelle d'entropie* $\mathcal{H}(\rho) := \int \rho \ln(\rho) dx$ *:*

$$\nabla^{L^2}_\rho \mathcal{H} = \ln(\rho) + 1$$

Remarque 3.

(1) $\nabla^{L^2} H$ *agit sur les dérivées* L^2 *qui sont d'intégrale nulle, donc* $\nabla^{L^2} H$ *est défini à une constante près. Dans le dernier exemple,* $\ln(\rho)$ *est aussi un gradient* L^2 *de* $\mathcal{H}$.
(2) *On pourrait considérer l'espace riemannien* $L^2(\mu)$ *plutôt que l'espace euclidien.*
   - *Pour une mesure absolument continue, on a simplement* $\nabla^{L^2(\rho)}_\rho H = \frac{1}{\rho} \nabla^{L^2(\mathcal{L})}_\rho H$.
   - *Pour une mesure empirique* $\mu = \sum_{i=1}^N a_i \delta_{x_i}$, *et* $H(\mu) = H(a_i, x_i)$, *on a* $(\nabla^{L^2(\mu)}_\mu H)_i = \frac{1}{a_i} \frac{\partial H}{\partial a_i}$

De même qu'il existe une formule pour lier les dérivées L^2 et Wasserstein, il existe également une formule pour les gradients : pour un chemin $\mu_t \in \mathcal{P}_a(M)$ (densité ρ_t) dérivable pour les deux structures, $\partial_t \rho_t = -\operatorname{div}(\rho_t \vec{v}_t)$ donc

$$\int \nabla^{L^2} H \partial_t \rho_t = -\int \nabla^{L^2} H \operatorname{div}(\rho_t \vec{v}_t)$$

On peut de nouveau utiliser le fait que $-\operatorname{div}$ est l'adjoint du gradient : $\int \nabla^{L^2} H \operatorname{div}(\rho_t \vec{v}_t) = -\int \nabla(\nabla^{L^2} H) \rho_t \vec{v}_t$ donc

$$\frac{d}{dt} H(\mu_t) = \int \nabla^{L^2} H \partial_t \rho_t = \int \nabla(\nabla^{L^2} H) \rho_t \vec{v}_t dx$$

On peut donc identifier

$$\nabla^W H = \nabla(\nabla^{L^2} H)$$

Exemple 4. *Si* $H(\mu) = \int f d\mu$, *on a vu* $\nabla^{L^2} H = f$, *donc* $\nabla^W H = \nabla f$

2.4. **Flots gradients Wasserstein.** On a pu identifier le gradient Wasserstein, on peut donc écrire l'équation de flot gradient

$$\left(\frac{d}{dt}\mu_t\right)_W = -\nabla^W_{\mu_t} H$$

Rappelez-vous que pour écrire $\left(\frac{d}{dt}\mu_t\right)_W = \vec{v}_t$ en eulerien pour des mesures μ_t absolument continues de densité ρ_t, on avait utilisé l'équation de continuité

$$\partial_t \rho_t = -\operatorname{div}(\rho_t \vec{v}_t)$$

Il nous suffit de remplacer $\vec{v}_t$ par $-\nabla^W_{\mu_t} H = -\nabla(\nabla^{L^2} H)$

Proposition 5 (descente de gradient)**.** *Formellement, la descente de gradient* $\left(\frac{d}{dt}\mu_t\right)_W = -\nabla^W_{\mu_t} H$ *dans* $(\mathcal{P}_2(M), W_2)$ *s'écrit en eulerien sous la forme*

$$\partial_t \rho_t = \operatorname{div}(\rho_t \nabla(\nabla^{L^2} H))$$

L'équation de diffusion, ou équation de la chaleur est l'exemple le plus frappant : dans l'exemple 3, on a vu que le gradient L^2 de l'entropie $\mathcal{H}(\rho) := \int \rho \ln(\rho) dx$ est

$$\nabla^{L^2}_\rho \mathcal{H} = \ln(\rho)(+1)$$

Donc le gradient Wasserstein de l'entropie est

$$\nabla^W_\rho \mathcal{H} = \nabla(\ln(\rho)) = \frac{\nabla \rho}{\rho}$$

D'où l'équation de continuité

$$\partial_t \rho_t = \operatorname{div}(\rho_t \frac{\nabla \rho_t}{\rho_t}) = \operatorname{div}(\nabla \rho_t)$$

On obtient donc l'équation de diffusion

$$\partial_t \rho = \Delta \rho$$

Flot hamiltoniens Wasserstein Une fois qu'on connait l'expression du gradient Wasserstein, on peut considérer toute sorte d'équations différentielles, par exemple au lieu de suivre les lignes de plus grande pente de H, on peut aussi suivre les lignes de niveau dans une dynamique hamiltonnienne par la formule

$$\left(\frac{d}{dt}\mu_t\right)_W = J\nabla^W_{\mu_t} H$$

où J est une matrice de rotation. On obtient une dynamique hamiltonnienne où l'énergie H est conservée. En eulerien, on a

$$\partial_t \rho_t = \operatorname{div}(\rho_t J \nabla(\nabla^{L^2} H))$$

Un exemple important est celui des équations semi-géostrophique. Une dernière remarque est que le champ de vecteur $J\nabla(\nabla^{L^2} H)$ est à divergence nulle (rotation d'un gradient), ce qui est très pratique.

Ces deux premières parties ont présenté un résultat classique qui éclaire le lien entre transport optimal et entropie. C'est en lien avec mon sujet de thèse puisque je devrais étudier des généralisations du transport optimal et comment utiliser la pénalisation entropique dans ces conditions.

3. QUESTIONS SUR L'ENTROPIE

Dans cette dernière partie, je voudrais introduire des questions que je me pose sur un sujet directement lié au résultat précédent.[5] Le cadre est désormais $M = \mathbb{R}^d$

Motivation L'entropie donne une bonne indication du degré de concentration d'une mesure. Avec la convention $\mathcal{H}(\mu) = +\int \rho \ln(\rho) dx$ (où ρ est la densité de μ), une grande entropie correspond à une grande concentration, une petite entropie à une mesure très diffuse. Si telle densité a une entropie plus grande que telle autre, on peut légitimement dire qu'elle est moins diffuse. Il serait agréable de pouvoir comparer ainsi des mesures qui n'ont pas forcément absolument continues, et qui peuvent avoir une entropie infinie.

La dimension de Renyi, ou dimension d'information répond partiellement à cette intérrogation. Cette dimension peut s'introduire grâce au transport optimal : soit

$$W_\infty(\mu, \nu) = \inf_{X \sim \mu, Y \sim \nu} \sup(d(X, Y))$$

C'est une distance très naturelle sur les mesures : $W_\infty(\mu, \nu) \leq t$ signifie qu'on peut apporter la masse de μ en ν sans jamais faire de trajets de taille supérieure à t. Soit aussi

$$H_0(\mu) := \begin{cases} \sum a_i \ln(a_i) & si\ \mu = \sum a_i \delta_{x_i} \\ -\infty & sinon \end{cases}$$

C'est l'entropie de dimension 0.

Définition 6 (proposition)**.** Soit $\mu \in \mathcal{P}(\mathbb{R}^d)$ Soit $h(\mu, t) := \sup_{W_\infty(\mu,\nu) \leq t} H_0(\nu)$ alors

$$h(\mu, t) \sim d_\mu \ln(t) \quad (t \to 0)$$

. où $d_\mu \in [0, d]$ est un nombre appelé *dimension de Renyi* de μ

Le nombre d_μ donne une échelle de diffusivité. De plus, si μ et ν sont de dimension entière $d_\mu = d_\nu \in \mathbb{N}$, on peut tenter de les comparer plus finement grâce à l'entropie de dimension d_μ : Par exemple si μ est concentrée sur un sous-espace V de dimension d_μ et absolument continue par rapport à la mesure de Lebesgue de dimension d_μ sur ce plan, notée dx^{d_μ}, alors $d\mu = \rho dx^{d_\mu}$ et on peut calculer

$$H_{d_\mu}(\mu) = \int_V \rho \ln(\rho) dx^{d_\mu}$$

Mais ce n'est pas toujours le cas, comme illustré par la figure 7

Minimiser H_d plutôt que maximiser H_0. Comme le résultat précédent sur l'équation de diffusion semble indiquer qu'il existe un lien spécial entre W_2 et l'entropie de dimension d, on peut tenter d'utiliser la même stratégie "en partant du haut" : en notant H_d la d-entropie des mesures absolument continues par rapport à $\mathcal{L}$ Soit

$$H(\mu, t) := \inf_{W_2(\mu,\nu) \leq t} H_d(\nu)$$

Comme l'équation de la chaleur est le flot gradient de l'entropie, il est naturel d'utiliser ρ_t solution de $\begin{cases} \partial_t \rho_t = \Delta \rho_t \\ \\ \rho_t \to \mu \quad (t \to 0) \end{cases}$ comme candidat pour minimiser $H_d(\rho)$ à $W_2(\rho, \mu) = W_2(\rho_t, \mu)$.

[5] Il est probable que la réponse à ces questions existe quelque part dans la littérature

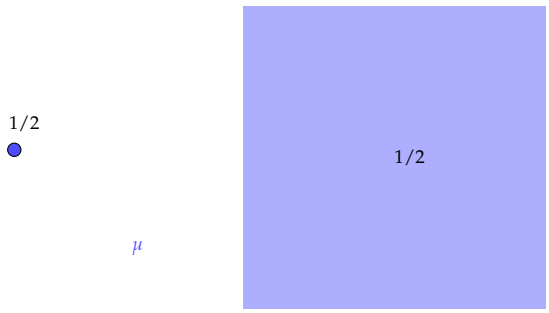

FIGURE 7. La mesure μ, à moitié dirac et à moitié absolument continue est de dimension de Renyi égale à 1. Mais il est dur de calculer sa 1-entropie

Solutions de l'équation de diffusion L'équation de diffusion est très étudiée, et sa solution est connue : la distribution gaussienne $p_t(x) = \frac{1}{(2\pi t^2)^{d/2}} e^{\frac{-\|x\|^2}{2t^2}}$ vérifie l'équation $(\partial_t - \Delta)(p_t) = 0$ et $p_t \to \delta_0 \quad (t \to 0)$. On appelle cette distribution le *noyau de la chaleur*[a]. si ρ_0 est la condition initiale, ρ_t définie par convolution avec p_t $\rho_t := p_t * \rho_0$ vérifie naturellement

$$(\partial_t - \Delta)(\rho_t) = (\partial_t - \Delta)(p_t * \rho_0) = (\partial_t - \Delta)(p_t) * \rho_0 = 0 * \rho_0 = 0$$

Et

$$\lim_{t\to 0} \rho_t = \lim_{t\to 0} (p_t) * \rho_0 = \delta_0 * \rho_0 = \rho_0$$

[a] l'équation de diffusion est aussi appelée équation de la chaleur

L'exemple élémentaire. Pour $\mu = \delta_0$, on prend directement le noyau de la chaleur p_t (voir encadré). $H_d(p_t) = -d\ln(\sqrt{2\pi e}) - d\ln(t)$ et $W_2^2(p_t, \delta_0) = dt^2$, donc

$$H(\delta_0, \sqrt{dt^2}) \leq -d\ln(\sqrt{2\pi e}) - d\ln(t)$$

et en renormalisant t par $\sqrt{d}$, on obtient :

$$H(\delta_0, t) \leq -d\ln(t/d) - d\ln(\sqrt{2\pi e})$$

Si bien que

$$H(\delta_0, t) \sim -d\ln(t) \quad (t \to 0)$$

On peut utiliser la tensorisation pour traiter des exemples un peu plus compliqués que δ_0 :

Tensorisation, $\mathcal{H}$ et W_2 L'entropie et la distance de Wasserstein se comportent très bien par tensorisation : Soit a,b,c et d quatres mesures de probabilité sur des espaces $\mathbb{R}^k$ de dimensions possiblement différentes. On peut utiliser la formulation probabiliste

$$W_2^2(a\otimes b, c\otimes d) = \sup_{\substack{(X_1,X_2,Y_1,Y_2)\\(X_1,X_2)\sim a\otimes b\\(Y_1,Y_2)\sim c\otimes d}} \mathbb{E}[\|X_1-Y_1\|^2+\|X_2-Y_2\|^2]$$

pour obtenir

$$W_2^2(a\otimes b, c\otimes d) = W_2^2(a,c)+W_2^2(b,d)$$

On a aussi, en notant sans distinction $\mathcal{H}$ l'entropie de toute dimension,

$$\mathcal{H}(a\otimes b) = \int a(x)b(y)\ln(a(x)b(y))dxdy$$

et en utilisant simplement $\ln(ab)=\ln(a)+\ln(b)$, on obtient

$$\mathcal{H}(a\otimes b) = \int a\ln(a)(\int bdy)dx + \int b\ln(b)(\int adx)dy$$

et comme a et b sont d'intégrale 1

$$\mathcal{H}(a\otimes b) = \mathcal{H}(a)+\mathcal{H}(b)$$

Questions. Avec cette méthode, et en considérant des exemples simples, on obtient les questions suivantes :

(1) Existe-t-il une constante c_μ
$$H(\mu,t)\sim -c_\mu\ln(t)\quad (t\to 0)$$
(2) c_μ est elle la codimension de μ, ie $c_\mu+d_\mu=d$?
(3) A-t-on $H(\mu,t)=-c_\mu\ln(t/c_\mu)+\ln(\sqrt{2\pi e})+f(\mu)+o(1)\quad (t\to 0)$ pour une fonction f ?
(4) A-t-on $h(\mu,t)=d_\mu\ln(t)+g(\mu)+o(1)\quad (t\to 0)$ pour une fonction g ?
(5) A-t-on $f=g$?
(6) A-t-on $f(\mu)=H_{d_\mu}(\mu)$ pour μ concentré sur un sous-espace de dimension d_μ ?

Conclusion

J'espère que ce document vous aura permis de découvrir mon domaine de recherche. Je ne pense pas utile de donner de référence d'articles précis, sauf tout de même pour celui de Jordan Kinderlehrer Otto qui a marqué la découverte des flots gradients Wassestein [2]. Si vous êtes intéressé par une présentation plus exacte et rigoureuse des différents pans du domaine, je vous propose ici quelques livres. D'abord et avant tout le premier livre de Villani, [5], pour une présentation complète et géométrique du transport optimal et des liens avec les équations différentielles. Pour une présentation moins géométrique mais bien plus précise sur les autres applications, le livre de Santambrogio [4] est très bien également. Tout comme [1] pour ceux qui s'intéressent aux liens avec l'économie. En ce qui concerne le transport optimal computationnel, le livre [3] est également intéressant.

References

[1] A. Galichon. *Optimal Transport Methods in Economics*. Princeton University Press, 1 edition, 2016.
[2] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.
[3] G. Peyré and M. Cuturi. Computational Optimal Transport. *ArXiv e-prints*, Mar. 2018.
[4] F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing, 2015.
[5] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.